Psychometrika

A JOURNAL DEVOTED TO THE DEVELOPMENT OF PSYCHOLOGY AS A QUANTITATIVE RATIONAL SCIENCE

THE PSYCHOMETRIC SOCIETY · ORGANIZED IN 1935

VOLUME 10
NUMBER 3
SEPTEMBER
1945

PSYCHOMETRIKA, the official journal of the Psychometric Society, is devoted to the development of psychology as a quantitative rational science. Issued four times a year, on March 15, June 15, September 15, and December 15.

SEPTEMBER 1945, VOLUME 10, NUMBER 3

Printed for the Psychometric Society at 23 West Colorado Avenue, Colorado Springs, Colorado. Entered as second class matter, September 17, 1940, at the Post Office of Colorado Springs, Colorado, under the act of March 3, 1879. Editorial Office, College Entrance Examination Board, Princeton, New Jersey.

Subscription Price:

To non-members, the subscription price is $5.00 per volume of four issues. Members of the Psychometric Society pay annual dues of $5.00, of which $4.50 is in payment of a subscription to *Psychometrika*. Student members of the Psychometric Society pay annual dues of $3.00, of which $2.70 is in payment for the journal.

The subscription price to libraries and other institutions is $10.00 per year; this price includes one extra copy of each issue.

Applications for membership and student membership in the Psychometric Society should be sent to

THELMA GWINN THURSTONE
Chairman of the Membership Committee
University of Chicago
Chicago, Illinois

Membership dues are payable to

IRVING D. LORGE
Treasurer of the Psychometric Society
425 West 123rd Street
New York, New York

Individual and library subscriptions are payable to, and change of address should be sent to

IRVING D. LORGE
Treasurer of the Psychometric Corporation
425 West 123rd Street
New York, New York

Articles on the following subjects are published in *Psychometrika*:

(1) the development of quantitative rationale for the solution of psychological problems;
(2) general theoretical articles on quantitative methodology in the social and biological sciences;
(3) new mathematical and statistical techniques for the evaluation of psychological data;
(4) aids in the application of statistical techniques, such as monographs, tables, work-sheet layouts, forms, and apparatus;
(5) critiques or reviews of significant studies involving the use of quantitative techniques.

(*Continued on the back inside cover page*)

Psychometrika

CONTENTS

VOLUME TEN SEPTEMBER 1945 NUMBER THREE

PSYCHOMETRIKA—VOL. 10, NO. 3
SEPTEMBER, 1945

THE EFFECTS OF SELECTION IN FACTOR ANALYSIS

L. L. THURSTONE
THE UNIVERSITY OF CHICAGO

Factorial results are affected by selection of subjects and by selection of tests. It is shown that the addition of one or more tests which are linear combinations of tests already in a battery causes the addition of one or more incidental factors. If the given test battery reveals a simple structure, the addition of tests which are linear combinations of the given tests leaves the structure unaffected unless the number of incidental factors is so large that the common factors become indeterminate.

1. *Types of Selection*

When a factor analysis has been completed for a test battery on one group of subjects, the administration of the same battery to a differently selected group of subjects will give a different factorial result if the second group is selected by any criterion in the common factors. If a simple structure was found for the first group of subjects, the same structure will be found for the second group, but the correlations between the primary factors will be different. Incidental factors are added to the common factors under certain conditions of selection of the subjects. It will be shown that under these various conditions of selection of tests and subjects, a simple structure remains invariant. It will also be shown that the physical interpretation of each primary factor remains the same for wide variations in selection of both subjects and tests but that their intercorrelations are altered by selection of subjects. It can readily be seen that these considerations are of fundamental significance for factor analysis as a scientific method.

These principles will be illustrated by numerical examples with the box problem. The box dimensions x, y, z, may have low correlations in one group of boxes but these correlations may be quite different if the analysis is repeated on another group of boxes so selected that they are all, say, relatively low and squatty. Then the x-variance is reduced and the correlations may be altered. Of course, these altered correlations of the primary factors do not deny the identity of the parameters length, height, and width, for describing the box measurements in both box populations. The factors are the same and their physical interpretations are the same in both populations even though their intercorrelations alter from one experimental population to another. This simple principle will be illustrated in numerical examples.

2. Experimental Dependence

One of the most important sources of incidental factors which are extraneous to the purposes of a factorial study is what we have called *experimental dependence*. In earlier studies with the factorial methods this source of factors was not recognized, but for a number of years this principle has been applied in the preparation of test batteries in the Psychometric Laboratory at the University of Chicago.* In *Table 1* we have the factor pattern for a set of five tests (1, 2, 3, 4, 5) with two common factors *A* and *B*. The reduced correlation matrix of this battery would be of rank 2, corresponding to the number of common factors. For each of these five tests there is also indicated in the factor pattern a unique variance. Now suppose that the performance from which the score of test 1 was derived is also used to derive another score 1a. Then we have two test scores derived from the same performance, and they are quite likely to share some of the uniqueness of that particular performance. This fact is indicated in the factor pattern by the two *x*'s in the third column. Tests 1 and 1a share the same uniqueness. In addition, the two tests 1 and 1a may also have unique factors that are not shared, but these are not indicated in the factor pattern of *Table 1*.

A simple example of this manner of constructing a test battery would be to record speed and accuracy as two scores from the same test performance in the attempt to cover all aspects of the domain that is being investigated. The new factor pattern, after adding test 1a, has three common factors. The third common factor has saturations in only 1 and 1a. It should be noted that the rank of the reduced correlation matrix has been raised from 2 to 3 by this manner of assembling the test battery. In factoring the test battery it would be necessary to extract one additional factor from the correlation matrix before the residuals vanish. After rotation of axes a third common factor would be found on which only two tests have significant saturations, namely, 1 and 1a. Upon examination of their nature it would be found that these two tests derive from the same test performance. Such a factor on which two variant scores of the same performance have saturations is called a *doublet factor* and it is ignored in interpretation. It rarely gives any leverage on the prob-

* The importance of exoerimental dependence in factor analysis was first pointed out by Ledyard Tucker.

lem of interpreting what the factor might represent except that it is something concerned with that test performance—a fact that was known before the factoring was begun. For this reason it is best to avoid inserting in a test battery two or more measures that are taken from the same test performance.

If we push this effect to an extreme by adding a duplicate form of each test to the battery, or by using each test and a parallel form of it, then we have a battery of ten tests instead of the five tests with which we started. If there is a unique variance for each test, then the factor pattern of *Table 1* is altered by introducing five new common factors so that there are seven common factors in all, the two factors common to the five original tests and the five unique variances which have now become common factors. It is doubtful whether a useful resolution of the correlation matrix into common factors is then possible. The unique variances have been made into common variances by the inclusion of a parallel form of each test and the common factors that might be significant would be obscured by the addition of five common factors that may make the factor problem insoluble.

It is good practice to inspect a test battery for experimental dependence before factoring is begun. One of the most common forms of experimental dependence is that in which some kind of total score is added to a test battery to be factored. A mental age, for example, is essentially a summation score. It is known before the analysis is begun that the total score will be represented by a test vector that is in the middle of the configuration and that, hence, it contributes nothing to the identification of the primary vectors which are at the corners or intersections of the configuration. Nor does it contribute to the identification of a second-order general factor which is defined by the correlations of the primaries. However, a total score may be found to have higher saturation on the second-order general factor than any one of the primaries. The best procedure is to investigate the structure of the domain without the summation score and to investigate the characteristics of the summation separately.

3. *Linearly Dependent Tests*

The summation score is an example of both experimental dependence and linear dependence of scores. If we add to a test battery a new test score which is merely some linear combination of test scores from the battery, we have linear dependence, which generally introduces a new common factor. An exception is that in which the tests that are combined have no unique variance—a situation that arises only rarely. The introduction of a new test which is a linear combination of tests already in the battery is represented by the factor pattern of *Table 2.* Here we represent five tests (1, 2, 3, 4, 5) with two common factors and five unique factors. If a new test were added by merely taking the sum of the test scores 1 and 2, we should have the test denoted (1 + 2). This test is indicated as participating in the two common factors A and B and in the uniqueness which adds two more common factors.

4. Univariate Selection and Multivariate Selection of Subjects

When a factor study has been done on a particular group of subjects with a given test battery, it is of fundamental interest to know how the factorial results would be altered if the analysis were repeated on another group of subjects differently selected. We are not here interested in whether one of the groups is in any way representative of a universe while the other group is not so representative. For our purposes the two groups are coordinate but they are differently selected so that they differ in homogeneity in one or more factors. The results for either group are assumed to be known and we want to be able to predict the factorial results when we know how the second group was selected in comparison with the first group.

Consider the selection of a second group of subjects so that they are more homogeneous, or less, in one of the tests in the battery. This test will be called the *selection test* and it will be denoted l. The variances of all the tests have been arbitrarily reduced to unity by standardizing the scores for the first group so that in that group $\sigma_j = 1$ and, in particular, so that $\sigma_l = 1$. Let the new group be selected so as to have a different standard deviation in the selection test l, and let it be denoted $s_l \neq 1$. The new group of subjects will have different standard deviations in all the tests that have correlation with the selection test l. These new standard deviations will be denoted s_j. The problem is then to estimate the new test correlations and the new factorial results for the second group. The selection of a new group on the basis of their homogeneity in one of the tests has been called *univariate selection* by Godfrey Thomson, who has investigated this problem and whose results will be summarized here. Univariate selection can be either *complete* or *partial.* If the new group of subjects is so selected that all of the subjects have the same score in the selection test l ($\sigma_l = 0$), then we have *complete univariate selection.* If the new group is selected so that $\sigma_l \neq 0$, then the selection will be called *partial univariate selection* as to the selection variable l. It will be shown that interesting differences in the factorial results are determined by univariate selection when $\sigma_l > 0$ and when $\sigma_l = 0$, which represents complete homogeneity in the selection test. Further, it will be shown that the factorial results are also affected by the communality of the selection test l so that the results differ according to whether $h^2_l = 1$ or $h^2_l < 1$.

When the selection of a second group of experimental subjects is determined by more than one selection test we have *multivariate selection.* The principles of multivariate selection can be illustrated in terms of two selection tests, denoted l and m. Such a situation can be analyzed as two *successive univariate selections.* In multivariate selection one is likely to be dealing with a composite selection test $c = f(1,m)$. This is the case of multivariate selection that has been studied by Godfrey Thomson and by Ledermann. Here also, the factorial results are affected by complete and partial selection in the composite criterion of selection and by the communalities of the selection variables. The theory of univariate and multivariate selection will be described and numerically illustrated, and a comparison will be made of the psychological interpretations of the resulting factors.

5. *Geometrical Representation*

Before writing the vector equations which represent the effects of univariate selection on the standard deviations and on the test correlations, we shall consider the geometrical interpretation of selection in terms of the test configuration. In *Figure 1* we have a set of six unit test vectors in a configuration of two dimensions. At the start the problem will be simplified by letting all six of these test vectors be of unit communality in the space of the two common factors so that the diagram represents the total test space as well. The standard deviations of the tests are represented by the unit length of the test vectors which correspond to the fact that $\sigma_j = 1$ for all the tests. The test correlations are represented by the scalar products of pairs of test vectors. In n dimensions we should have a test configuration which is spherical in that the dispersions have all been normalized by reducing the scores to standard scores. The correlations are unaffected by this standardization.

Now let another group of experimental subjects be selected so that they show less variability in the direction of a selection test L. The effect will be to alter the shape of the test configuration so that it is narrower in the direction of the selection test. In fact, the configuration will have an ellipsoidal shape so that the configuration of two dimensions in *Figure 1* will be an ellipse determined by the termini of the test vectors. Since the new group is selected so as to be more homogeneous in test L, it follows that the components in this direction of all the other tests will be affected proportionally. If the factor of test L is reduced in dispersion, the same factor is also reduced proportionally in all the other tests in which this factor enters as a component.

Figure 1 has been drawn to represent a reduction of one half in the dispersion of test L. The L-component of the other tests is then also reduced by one half so that the new test vectors are depressed in the direction parallel to test L as shown in the diagram. The termini of the new test vectors now define an ellipse instead of the circle. It should be noted that tests A and E are unaffected by the selection because these tests are orthogonal to (uncorrelated with) the selection test.

The scalar product of the unit vectors L and D is the given correlation r_{LD}. This correlation is also the cosine of the angle LOD and it is the projection of the vector D on the vector L. This is the distance DX. It is the component of D in the direction of L. Reducing this component by one half we have the new test vector d as shown. The length of this vector is the standard deviation of test D in the selected group. The new vector for test L is l, which is collinear with L but only half as long. The cross product ld is the covariance of tests L and D in the new group.

If each of these shortened vectors, l and d, is extended to unit length we have vectors L' and D'. It should be noted that the vectors L and L' are identical whereas D and D' are not the same vectors. The correlation between two tests, such as C and D, is changed by the selection to a value which is the scalar product of the new vectors C' and D'.

6. The Correlations After Selection

The effect of univariate selection on the test correlations can be shown vectorially. In *Figure 2* let L be a unit vector representing the selection test in the total test space. Let J be a unit vector representing any other test in the battery. The components of J may be considered to be OX, which is orthogonal to L, *and* XJ, which is collinear with L. When selection is made on the basis of dispersion in L, the component OX of J will be unaffected whereas the component XJ will be affected.

Let the selection alter the dispersion of L so that its standard deviation is altered from $\sigma_l = 1$ to $s_l = p$. The test L is then represented by the test vector l, whose length is $s_l = p$. The new test vector l can be written in the form

$$l = L - tL, \tag{1}$$

where

$$t \equiv 1 - p. \tag{2}$$

The component XJ of J is the projection of J on L, and it is also the correlation r_{jl}. Since the effect of selection on the test vector J is proportional to its component collinear with L, we have

$$J_0 = J - t\, r_{jl}\, L\,, \tag{3}$$

where J_0 is the new vector for test J. Its new standard deviation is the length of the vector J_0. Squaring (3), we have

$$J^2_0 = (J - t\, r_{jl}\, L)^2 \tag{4}$$

or

$$J^2_0 = s^2_j = J^2 - 2\, t\, r_{jl}\, L\, J + t^2\, r^2_{jl}\, L^2. \tag{5}$$

The scalar product JL is the correlation r_{jl}, and the scalar product $L^2 = J^2 = 1$. Hence

$$s^2_j = 1 - 2\, t\, r^2_{jl} + t^2\, r^2_{jl}\,, \tag{6}$$

which becomes

$$s^2_j = 1 - r^2_{jl}\,(2\, t - t^2). \tag{7}$$

For convenience, let $q^2 \equiv 1 - p^2$. From (2) we then have

$$2\, t - t^2 = q^2, \tag{8}$$

so that the new variance s^2_j can be written as

$$s^2_j = 1 - q^2\, r^2_{jl}\,, \tag{9}$$

by which the new standard deviation becomes

$$s_j = \sqrt{1 - q^2\, r^2_{jl}}\,. \tag{10}$$

The numerical values of s_j can then be computed, since the correlations r_{jl} and the constants p and q are assumed to be known.

By analogy with (3) we can write the corresponding vector equation for test K, and we have then

$$K_0 = K - t\, r_{jl}\, L\,. \tag{11}$$

The new covariance c_{jk} can be expressed as the scalar product J_0K_0, which by (3) and (11) is

$$c_{jk} = J_0K_0 = (J - t\, r_{jl}\, L)\,(K - t\, r_{jl}\, L), \tag{12}$$

and this becomes

$$c_{jk} = J\, K - t\, r_{kl}\, J\, L - t\, r_{jl}\, K\, L + t^2\, r_{jl}\, r_{kl}\, L^2\,. \tag{13}$$

Recalling that the scalar products $JK = r_{jk}$, $JL = r_{jl}$, $KL = r_{kl}$, and $L^2 = 1$, we get

$$c_{jk} = r_{jk} - t\, r_{jl}\, r_{kl} - t\, r_{jl}\, r_{kl} + t^2\, r_{jl}\, r_{kl}\,. \tag{14}$$

Collecting terms,

$$c_{jk} = r_{jk} - r_{jl}\, r_{kl}\,(2\,t - t^2), \tag{15}$$

which by (8) becomes

$$c_{jk} = r_{jk} - q^2\, r_{jl}\, r_{kl}\,. \tag{16}$$

This equation enables us to compute the new covariances to be expected by selection on test L.

The new correlations can be obtained by the stretching factor which merely reduces the new scores to standard scores. Vectorially this stretching factor is represented by normalizing the new test vectors J_0 and K_0 to unit length. Since the lengths of these vectors are known to be their standard deviations as given by equation (10) we have for the new correlations

$${}_s r_{jk} = \frac{c_{jk}}{s_j\, s_k}, \tag{17}$$

or, in more complete form,

$${}_s r_{jk} = \frac{r_{jk} - q^2\, r_{jl}\, r_{kl}}{s_j\, s_k}, \tag{18}$$

where ${}_s r_{jk}$ denotes the test correlations after selection. This is Godfrey Thomson's equation for the new correlations after selection.* In computing the new correlations ${}_s r_{jk}$, it should be recalled that $r_{jj} = r_{ll} = 1$ because the present problem is concerned with the total variances of the tests. It is not limited to the common factors. The diagonals of the given correlations r_{jk} are unity.

In general the dimensionality of the test configuration in the total test space is equal to the number of tests n. When the test scores are normalized, each test vector is of unit length in the total test space and the surface determined by the test vector termini is a sphere in n dimensions. After selection on one of the tests so as to reduce the variance of the selection test L, we have an ellipsoid determined by the test vector termini. If the selection on test L is complete so that all subjects in the new group have the same score in test L, then we should expect to lose one dimension. The surface

* Thomson, Godfrey H. "The influence of univariate selection on the factorial analysis of ability. *Brit. J. Psych.* (Gen. Sec.) 1938, 28, Part 4.

Ledermann, Walter. Note on Professor Godfrey H. Thomson's article "The influence of univariate selection on the factorial analysis of ability." *Brit. J. Psychol.* (Gen. Sec.) 1938, 29, Part 1.

Thomson, Godfrey H. and Ledermann, Walter. The influence of multivariate selection on the factorial analysis of ability. *Brit. J. Psychol.* (Gen Sec.) 1939, 29, Part 3.

Thomson, Godfrey H. *The factorial analysis of human ability,* Chapters 11 and 12.

determined by the test vector termini would then be a sphere of dimensionality $(n-1)$. If $n=3$ the spherical surface becomes a round pancake when p is small and a plane when p is zero.

When $s_l = p = 0$ so that $t = 1$, the selection on test L is complete. The dimensionality is then reduced to $(n - 1)$ and equation (18) becomes

$$_sr_{jk} = r_{jk \cdot l} = \frac{r_{jk} - r_{jl}\, r_{kl}}{\sqrt{1-r^2_{jl}}\ \sqrt{1-r^2_{kl}}}, \tag{19}$$

which is the familiar partial correlation formula. If the three-dimensional configuration has been represented on the surface of a sphere and if the selection is complete on test L, the resulting two-dimensional configuration can be visualized by looking at the spherical model with test vector L in the line of regard. The resulting two-dimensional configuration is then directly seen. The partial correlations would then be represented by the scalar products of pairs of these vectors after they have been normalized in the two dimensions that would be seen on a photograph of the sphere with test L at the center of the picture.

It has been pointed out by Godfrey Thomson that the well-known Otis-Kelley formula for the effect of restricted selection on the correlation coefficient is a special case of equation (18). In this case $s_j = s_k$ because both variables are then assumed to be subject to the same restricted selection. Equation (18) then becomes

$$_sr_{jk} = \frac{r_{jk} - q^2\, r_{jl}\, r_{kl}}{s^2_j}. \tag{20}$$

By (10) we have, when $s_j = s_k$,

$$1 - q^2\, r^2_{jl} = 1 - q^2\, r^2_{kl}, \tag{21}$$

so that

$$r_{jl} = r_{kl} \tag{22}$$

and hence (20) becomes

$$_sr_{jk} = \frac{r_{jk} - q^2\, r^2_{jl}}{s^2_j}. \tag{23}$$

From (9) we have

$$q^2\, r^2_{jl} = 1 - s^2_j, \tag{24}$$

so that (23) becomes

$$_sr_{jk} = \frac{r_{jk} - (1 - s^2_j)}{s^2_j} \tag{25}$$

and hence

$$ {}_sr_{jk}\, s^2_j = r_{jk} - 1 + s^2_j \,, \tag{26} $$

so that

$$ s^2_j (1 - {}_sr_{jk}) = 1 - r_{jk} \,, \tag{27} $$

from which we have

$$ \frac{1 - r_{jk}}{1 - {}_sr_{jk}} = s^2_j = \frac{s^2_j}{\sigma^2_j} , \tag{28} $$

which is the Otis-Kelley formula. These estimates of the effects of selection are of course in the nature of expected averages from which the selected samples would show random deviations.

7. The Communalities

So far we have considered the effect of selection on the dispersions and correlations which imply the total test space. When the communalities are known in one correlation matrix, the new communalities can be found by a slight adaptation of equation (18). The diagonal correlations are then written $r_{jj} = h^2_j$ instead of unity. Since the selection is assumed to be made by the test scores, it follows that the uniqueness of the selection test as well as its common factor variance must be involved in the result.

Writing equation (18) for the communalities, where $r_{jj} = h^2_j$, we have

$$ {}_sr_{jj} = {}_sh^2_j = \frac{h^2_j - q^2\, r^2_{jl}}{1 - q^2\, r^2_{jl}} \qquad (j \neq l) \,, \tag{29} $$

where the denominator is written by equation (10). It should be noted that equation (29) applies to the computation of all the communalities except for the selection test. Hence the restriction for equation (29) that $j \neq l$.

When $j = l$, we can write this equation in the modified form

$$ {}_sr_{ll} = {}_sh^2_l = \frac{h^2_l - q^2 \cdot r_{ll} \cdot h^2_l}{1 - q^2\, r_{ll}\, h^2_l} , \tag{30} $$

where one of the correlations r_{ll} is unity and the other is h^2_l. Then

$$ {}_sh^2_l = \frac{h^2_l (1 - q^2)}{1 - q^2\, h^2_l} = \frac{h^2_l\, p^2}{1 - q^2\, h^2_l} , \tag{31} $$

by which the new communality of the selection test can be computed. Equation (31) may be derived vectorially as follows.

Let L in *Figure 3* be the selection test vector in the total test space and let L_c and L_u be the common factor component and the unique component of L so that $L = L_c + L_u$. By selection the configuration of test vectors is first contracted by the proportion p in the direction of L and then normalized. Contracting L to pL and then normalizing reproduces L again. Hence the selection vector L is not altered by selection except, of course, when $p = 0$ in which case pl becomes a null vector and the configuration is reduced to dimensionality $(r - 1)$.

The length of the common factor component L_c is h_l and the projection of L_c on L is pL, whose length is h^2_l. The correction vector for L_c which is introduced by selection is therefore $-th^2_l L$ as shown in *Figure 3*. The altered vector L_c can now be written as

$$_sL'_c = L_c - th^2_l L \, , \tag{32}$$

where $_sL'_c$ represents the distortion of L_c which is introduced by compressing the configuration in the direction of L before the vectors are finally normalized again.

In order to find the length of $_sL'_c$ we write

$$(_sL'_c)^2 = L^2_c - 2\,t\,h^2_l\,L_c\,L + t^2\,h^4_l\,L^2. \tag{33}$$

Since $L^2 = 1$, $L^2_c = h^2_l$, and since the scalar product $L_c\,L = h^2_l$, equation (33) becomes

$$(_sL'_c)^2 = h^2_l - 2\,t\,h^4_l + t^2\,h^4_l \, , \tag{34}$$

which by (8) reduces to

$$(_sL'_c)^2 = h^2_l\,(1 - q^2\,h^2_l). \tag{35}$$

The scalar product of $_sL'_c$ and L in *Figure 3* can be written either as a product from equation (32) or as the product of the lengths of L and $_sL'_c$ and the cosine of their angular separation. The length of L is unity, the length of $_sL'_c$ is $h_l\sqrt{1 - q^2\,h^2_l}$, and since the cosine of the angular separation is also the projection of L on the new common factor space, namely, $_sh_l$, we have

$$_sL'_c\,L = {_sh_l} \cdot h_l\sqrt{1 - q^2\,h^2_l}\,. \tag{36}$$

The same scalar product can be written by equation (32)

$$_sL'_c\,L = L\,L_c - t\,h^2_l\,L^2 \tag{37}$$

which becomes

$$_sL'_c\,L = h^2_l - t\,h^2_l \tag{38}$$

or

$$_sL'_c\,L = h^2_l\,p\,. \tag{39}$$

Equating (36) and (39) we get

$$h^2_l\, p = {}_s h_l\, h_l \sqrt{1 - q^2\, h^2_l} \tag{40}$$

and writing this explicitly for the new communality of the selection test we have

$${}_s h^2_l = \frac{h^2_l\, p^2}{1 - q^2\, h^2_l} \tag{31}$$

as previously written. By this equation the new communality of the selection test can be computed.

In solving a complete problem of this kind it is convenient to use Godfrey Thomson's equation (18) for all of the new correlations including the communalities except that of the selection test for which equation (31) applies.

8. Computational Sequence

The problem of determining the effects of univariate selection starts with the correlation matrix and the communalities as well as the degree of selection on the selection test, which is denoted L with subscript 1. It is assumed that the given scores have been standardized so that $\sigma_j = 1$ for all tests. The selection of a new group is such that the standard deviation of the selection test is changed from the given value $\sigma_l = 1$ to a new value s_l which is known. The ratio of the standard deviations is

$$\frac{s_l}{\sigma_l} = s_l = p\,,$$

where p is a parameter that determines the new factorial results. The following convenient constants are then computed: p^2, $q^2 = 1 - p^2$, and q.

The given correlation matrix is written with unity in the diagonals. Equation (18) can be written in the form

$$R_s = MCM = M(R - UU')M\,, \tag{41}$$

where R_s denotes the expected correlation matrix after univariate selection. The matrix U is a column vector with elements $u_j = q\, r_{jl}$, and the matrix M is a diagonal matrix with elements $d_j = 1/s_j$. Equation (41) indicates the computational order. One may compute first the column vector U, then the covariance matrix $C = UU'$ of order $n \times n$, and finally R_s. The correlation matrix R_s is, of course, symmetric and it also has unity in the diagonals.

When the correlation matrix R_s has been computed, the communalities may be computed separately by equation (36), where it must be remembered that r_{ll} is written as h_l.

9. Examples of Univariate Selection

The principles of the preceding sections will be illustrated by a numerical example with a physical model. A variant of the box problem will be used. The factor matrix of *Table 4* represents ten measures of the box problem including the basic parameters height, width, and length, and such other box measurements as the diagonal of each face, the area of each face, and one complex measure, such as the volume. In order to illustrate the effects of selection the given example in *Table 4* is arranged to represent the situation in which the three basic parameters are uncorrelated. Further, in order to show the effect of selection on the communalities, these have been arranged in the factor matrix as markedly below unity, namely, .70 for all of the ten measures. The corresponding correlation matrix is shown in *Table 3*. These are the given data.

It will now be supposed that a new collection of boxes is measured which is more homogeneous in the measurement or test No. 1, so that its standard deviation is .60 in the new population instead of unity as in the given population. The problem is now to investigate the effects of such selection on the correlations and on the simple structure. The reduced standard deviation is the parameter p of *Table 5,* and from this value are determined the other parameters such as q, t, and the new standard deviations s_j of the other measures that are correlated with test No. 1. These are all listed in *Table 5.*

In the same table we have the detailed computations for the new standard deviations s_j and the reciprocal m_j. The computations of this table are determined by equations (9) and (10).

The new covariances after selection are determined by equation (16), which can also be written in matrix form as shown in *Table 6.* The entries in this table are the cross-products of the new scores whose standard deviations have been depressed. Hence these entries are not correlation coefficients. The diagonal entries are the new variances of the tests. The corresponding correlation coefficients are shown in *Table 7,* and these represent Godfrey Thomson's equation (18) with unit diagonals. They can be obtained from the covariances of *Table 6* by the stretching factors $m_j = 1/s_j$, which are the elements of the diagonal matrix $\boldsymbol{M}$. This matrix is identical with the table of covariances except for the fact that the standard deviations have been stretched to unity.

The new communalities are determined by the detailed calculations shown in *Table 8* from (36) in which r_{ll} is written as h_l. Finally, when the new correlation matrix with communalities in the diagonal cells is factored we get the new factor matrix $\boldsymbol{F}_s$ as shown in *Table 9*. Here the number of factors is again three as in the given correlation matrix, thus verifying Godfrey Thomson's theorem that the rank of the correlation matrix with communalities remains unaltered with univariate selection. However, an exception should be mentioned. In complete or total univariate selection on a factor or its equivalent, a test with perfect communality, the new standard deviation of the selection test or factor is zero and the rank of the new correlation matrix has then been reduced by one. In the present example, the rank of the new correlation matrix, after complete selection by a factor or by a perfect test, is reduced to 2. Several examples of this type will also be shown.

Perhaps the most important consideration in this problem is to ascertain what happens to the identification of the primary factors. In *Figure 4* we have the configuration of test vectors before and after selection on test 1. In this diagram the test vectors are represented by the method of extended vectors so that the three-dimensional configuration can be seen in a plane. The given positions of the test vectors are denoted by *a* and the new positions by *b* . A glance at this figure shows that the position of the selection variable 1 remains unaltered. So do also the tests which are uncorrelated with the selection test. These are tests 2, 3, 8, 9. These remain unaltered. The other test vectors, 4, 5, 6, 7, 10, move in a direction radially from the selection vector so that the triangular configuration is still retained. In the present example the actual dimensions of the triangle even remain unaltered. We see, then, that in this example of partial univariate selection the simple structure remains invariant. If two investigators started with the two correlation matrices, one with the first group of objects and the other with the specially selected group of objects, they would arrive at the same simple structure and they would identify the same primary factors or parameters. But their results would differ as to the correlations between the measurements with which they started. In the present example, the correlations between the primary factors would remain the same for the two groups of objects. Variations of these principles will be shown in additional examples.

The most important principle to be drawn from analyzing the effects of selection on factorial results is that under wide variations in the conditions of selection the simple structure is invariant so that the primary factors or parameters are the same. The correlations of primary factors are altered from one selected group to another. A simple example will serve to illustrate this principle, which is well known in other contexts. The correlations between height, weight, and intelligence can be made to take widely different values depending on the selection of the experimental group. If children of all ages and statures are included, then the correlation between stature and intelligence will be found to be appreciable. This correlation is not spurious. It is correct to say that there actually is a high correlation between intelligence and stature for any experimental population of children with a wide range in chronological age. The taller children are generally older and they score higher on the tests. If the experimental population is limited to point age, the correlation between stature and intelligence vanishes, or nearly so. The interpretation is not to deny the meaning of the variables or the factors. These are the same measurements of stature and test score for the two groups. It is the correlations between the factors that are altered by the selective conditions. In the box example, it will be seen that the correlations between height, width, and length of the boxes can be altered by selection of each collection of boxes, but the physical interpretation of the three parameters remains precisely the same in the several box collections. We dwell on this point in detail because it is the source of misunderstanding of the factorial methods in that some critics are inclined to deny the validity of the physical interpretation of factors merely because their correlations are altered by conditions of selection.*

We turn now to several other examples of univariate selection with the same box problem. The detailed computations will not be repeated because the principles have been illustrated with the first numerical example. The next few examples have been prepared to show the effects of univariate selection of *factors* as distinguished from selection by *tests*. The three primary factors for the box problem may be denoted P_1, P_2, P_3. These may be regarded as unit vectors determined by the first three measures, 1, 2, 3, in the battery. *Figure 5* has been drawn to show the effects of univariate selection on the factor P_1. In this set of diagrams, *Figure 5* to *Figure 10* inclusive, the selection variable is denoted S. The method of extended

* Godfrey Thomson, who has contributed fundamental theory on the problem of univariate and multivariate selection in relation to factor analysis, has expressed doubt as to whether the primary factors can be interpreted as basic and identifiable psychological processes. His reason is mainly that the correlations between the primary factors of a simple structure are altered and determined in part by the conditions of selection and, further, that incidental factors can be added by certain conditions of multivariate selection. Our interpretation is that the primary factors represent the same basic processes in different conditions of selection and that it is the correlations between these parameters that are altered rather than the fundamental meaning of the parameters themselves. The incidental factors can be classed with the residual factors which reflect the conditions of particular experiments. These extraneous factors do not show the invariant characteristics that have been shown for the more basic primary factors. In presenting Godfrey Thomson's theoretical work on the selection problem, we are giving a less pessimistic interpretation of the factorial results.

vectors has been used to show the entire three-dimensional configuration in a single diagram in the plane of the paper. The given positions of the test vectors are shown by their numerical identification 1, 2, 3, etc. The new positions are denoted $1x$, $2x$, $3x$, etc. This set of six diagrams was designed by Ledyard Tucker to show graphically the effects of various conditions of univariate selection on the test configuration.

Figure 5 shows the effect of partial selection on the factor P_1. Note that the position of the selection variable $S = 1 = 1x$ remains unaltered. The positions of the test vectors 2, 3, 8, 9, remain the same because they are uncorrelated with the factor P_1. The other test vectors move radially from the selection variable S as shown in the triangular configuration. The simple structure is unaltered. In the same figure we have a diagram showing the effect of total selection on P_1. This means that all of the boxes in the new box population have the same height. The individual box shapes will be determined by only two parameters, since one of the parameters has become a constant and does not affect the individual differences among the objects. Therefore we expect the factorial result after selection to be of rank 2. This is verified in the diagram which was plotted after making the computations that have been explained. The variable 1 disappears from the second analysis because it is a null vector. This is what happens occasionally when the communality of a variable turns out to be vanishingly small. It usually happens when the corresponding correlations are also small. The new configuration lies entirely in the base line of the figure which represents only two dimensions.

If an investigator were to make a factor analysis of such a collection of boxes, he would find only two factors. If he were to plot the resulting configuration he would get a simple structure as shown in *Figure 6*. Here we see that even though one of the three factors has been entirely eliminated from the individual differences and from the analysis, the same simple structure is identified for the remaining two factors. The interpretation of the two primary factors, width and length of boxes, is the same in the two box populations.

Figure 7 was drawn to show the effect of univariate selection on the composite variable $(P_2 + P_3)$. The first diagram shows the effect of partial selection on this composite variable, which is denoted S. The given configuration of ten vectors is shown on the inside triangle, which is identical with that of *Figure 5*. The effect of selection is to move the termini of the test vectors radially from the selection vector to the new positions denoted $1x$, $2x$, $3x$, etc. The variable $1 = 1x$ is unaltered because it is orthogonal to the composite variable S. The others move as shown by the radial lines from S. The new configuration shows the same simple structure as before with the same primary factors in the tests 1, 2, 3. Hence the interpretation of the primary factors would be the same as before selection. The correlations between the primary factors have been altered.

In the second diagram of *Figure 7* we have the effect of total selection on the composite variable $S = (P_2 + P_3)$. Here we start, as before, with the given triangular configuration of ten test vectors extended to the tangent plane. The new configuration, after selection on S, is of two dimensions as is to be expected after complete selection on a factor. The new configuration is indicated along the upper horizontal line of the diagram. Tests $2x$, $9x$ are orthogonal to test 1 and so are also tests $3x$ and $8x$. If an investigator were to start with the selected box population and proceed with a factor analysis, he would find only two factors in this case. The configuration that he would find is shown in *Figure 8,* where it will be seen that the simple structure is again the same as before for the two factors that remained in the system. The primary factors are again identified by test 1 and the combination of tests 2 and 3. The physical interpretation of the factor P_1 would be the same as before. The interpretation of factors P_2 and P_3 would be obscured because the investigator would supposedly not know that he was dealing with a freak collection of boxes which had been so selected that the sum of the width and length of each box was the same for all of them. In the first diagram of *Figure 7* it is seen that with partial selection this ambiguity does not arise.

We turn next to an even more stringent selective condition, namely, the composite variable $S = (P_1 + P_2 + P_3)$. This is the sum of the three box dimensions. In *Figure 9* the inside triangle shows the given configuration as before. The composite variable S is in the middle of the configuration at test 10. The new configuration, after partial selection on the composite variable S, is shown by the outside triangular configuration. The new test vectors all lie in the outside triangle and their locations are determined by the radial lines from the selection variable as before. It can be seen that the simple structure is here again retained after selection and that the primary factors are the same, namely, those which are determined by tests 1, 2, 3. Hence we conclude that partial selection on a linear combination of even all of the factors leaves the simple structure invariant as well as the physical interpretation of the primary factors.

However, if the boxes are assembled so as to satisfy the conditions of total selection on the composite variable S, then the rank is reduced to 2 and in *Figure 10* we have the plot that an investigator would make with the new configuration. Here it is seen that the simple structure has been destroyed so that the primary factors are no longer identifiable. We conclude, therefore, that conditions of selection can be made so extreme that the simple structure is destroyed. In such a population the primary factors cannot be determined. In this case we would be dealing with a set of boxes so selected that the sum of the three box dimensions was the same for every box in the collection. For partial selection, as shown in *Figure 9,* the simple structure remains invariant. Hence we conclude that it is the total selection on this composite variable which destroys the simple structure.

10. Multivariate Selection of Subjects

So far it has been assumed that the selection of subjects is determined by their homogeneity in a *single* criterion variable even though this variable is itself a linear combination of several factors. Selection of subjects can be made to depend on several variables with the restriction that the intercorrelations of these variables shall have certain prescribed values. In these cases it may not be possible to describe the correlations in the selected group as an alteration in the homogeneity of a single variable, not even a composite selection criterion. We then have *multivariate selection* of subjects. However, multivariate selection can be described in terms of successive univariate selection by the formulas already discussed provided that the successive variables are properly chosen. This procedure will be illustrated by a numerical example that has been used by Godfrey Thomson in his discussions of multivariate selection.*

In analyzing multivariate selection it is convenient to divide the variables into two groups, namely, those which are directly involved in the selection and those which are not directly involved. The former, by which the selection is determined, will be denoted by subscript j and the others will be denoted by subscript k. The given correlation matrix can then be sectioned in the following manner:

$$\begin{array}{|c|c|}
\hline
\boldsymbol{R}_{jj} & \boldsymbol{R}_{jk} \\
\hline
\boldsymbol{R}_{kj} & \boldsymbol{R}_{kk} \\
\hline
\end{array}$$

The correlations of the selection tests are shown in section $\boldsymbol{R}_{jj}$ before selection. The correlations of the same tests after selection may be

* Thomson, Godfrey. The factorial analysis of human ability, page 187.

symbolized by the matrix V_{jj} which shows the new correlations that are imposed by selection. The new correlation matrix can then be symbolized in sectioned form as follows:

$$\begin{array}{|c|c|}
\hline
V_{jj} & V_{jk} \\
\hline
V_{kj} & V_{kk} \\
\hline
\end{array}$$

which represents the correlations to be expected after selection. The diagonal elements of V_{jj} show the new variances of the selection tests and the side entries of V_{jj} show the covariances to be imposed by selection.

The problem is then to determine the expected correlations in the sections V_{kk} and V_{jk} which is the transpose of V_{kj}. A formal matrix solution to this problem has been written by A. C. Aitken and reported by Godfrey Thomson (p. 189).* The solution is as follows:

$$V_{jk} = V_{jj} R^{-1}{}_{jj} R_{jk} , \tag{42}$$

$$V_{kk} = R_{kk} - R_{kj} (R^{-1}{}_{jj} - R^{-1}{}_{jj} V_{jj} R^{-1}{}_{jj}) R_{jk} , \tag{43}$$

$$V_{kj} = V'_{jk} . \tag{44}$$

Godfrey Thomson's numerical example will be used for the present discussion to illustrate the analysis of multivariate selection in terms of successive univariate selection. His numerical example is reproduced in the correlations between the six numbered variables of *Table 11.* The columns L and M have been added and they will be described presently. The given correlation matrix is of order 6 × 6 and it is of unit rank. The corresponding one-column factor matrix is shown in *Table 10.* In order to illustrate multivariate selection Thomson imposes a change in the variances of tests 1 and 2 and also in the covariance c_{12}. The given section of the correlation matrix for these two variables is

$$R_{jj} = \begin{array}{c|c|c|}
 & 1 & 2 \\
\hline
1 & 1.00 & .72 \\
\hline
2 & .72 & 1.00 \\
\hline
\end{array}$$

where the unit diagonals represent the unit variances of the tests at the start and where the given correlation $r_{12} = .72$. This section of

* Ibid.

the correlation matrix is to be changed by selection to

$$V_{jj} = \begin{array}{c|c|c|} & 1 & 2 \\ \hline 1 & .36 & .30 \\ \hline 2 & .30 & .36 \\ \hline \end{array}$$

where the two test variances are depressed to .36 and the covariance to $c_{12} = .30$. The problem is now to determine the expected covariances in the rest of the 6 × 6 matrix and then to factor the corresponding correlation matrix. The direct method of Aitken can be used to solve the problem. It can also be represented as a form of successive univariate selection.

In *Figure 11* the two unit vectors T_1 and T_2 represent the two tests 1 and 2. They have been drawn so that the cosine of the angular separation is .72, which is the given correlation r_{12}. The relation between these two tests which is imposed by selection is shown by the two vectors T_1' and T_2' with lengths of .60 and a scalar product of .30 as required by the new matrix V_{jj}. The two configurations have been arranged in *Figure 11* to utilize the symmetry in this case to simplify the transformation from one to the other. The new orthogonal selection variables are drawn in the figure, namely, L and M.

The transformation from the given configuration to the new configuration can be expressed in terms of two successive univariate selections, first on M and then on L. The first selection on M is specified by the restriction that

$$p = \frac{.5744}{.9274} = .6193,$$

so that the new vectors become $(T_1 - A)$ and $(T_2 - A)$. Then follows a univariate selection on L with the value

$$p = \frac{.1732}{.3741} = .4630,$$

so that the new vectors become

$$T_1' = (T_1 - A - B), \tag{45}$$

$$T_2' = (T_2 - A + B), \tag{46}$$

which define the new configuration and the covariances.

The first step is to add two rows and columns to the given correlation matrix for L and M in *Table 11*. The unit vector M bisects the angle between T_1 and T_2, so that we can write

$$M = c_1(T_1 + T_2). \tag{47}$$

Hence

$$r_{jm} = c_1(r_{j1} + r_{j2}), \tag{48}$$

where

$$c_1 = \frac{1}{\sqrt{2(1 + r_{12})}}. \tag{49}$$

In the same manner the unit vector L is a linear combination of T_1 and T_2, orthogonal to M, so that

$$L = c_2(T_1 - T_2), \tag{50}$$

and hence

$$r_{jl} = c_2(r_{j1} - r_{j2}), \tag{51}$$

where

$$c_2 = \frac{1}{\sqrt{2(1 - r_{12})}}. \tag{52}$$

Having thus determined the correlations of the six tests with L and M in *Table 11* we determine the new covariances after selection on M. These covariances are shown in *Table 12.* The covariances after further selection on L are then computed in the same manner and these are listed in *Table 13.* It is not necessary here to express *Table 12* in the form of correlation coefficients.

In *Table 14* are recorded the correlation coefficients corresponding to the covariances of *Table 13,* and in *Table 15* we have the final factor matrix after selection, which is also represented in *Figure 12.* It will be seen that the final factor matrix is of unit rank except for the doublet factor in the selection tests 1 and 2. The nature of this doublet factor is explained in *Table 2* referred to earlier in this article. The additional test is there a linear combination of the first two tests of the battery. Here the same effect is shown because the two unique factors have been introduced into the common factors to become an incidental common factor. The multivariate selection with the imposition that the new correlation r_{12} shall have an arbitrarily chosen value of $r_{12} = .83$ introduces an incidental factor, namely, the doublet factor II in *Table 15.* This introduction of an incidental factor into the common factors has not disturbed the single common factor I with which the problem started in *Tables 10* and *11.* The saturations of the factor I have been altered but not its identity.

11. Summary

Since factor analysis starts with a set of measurements for each individual member of an experimental group, it is evidently of fundamental significance to know how the factorial results are affected by the manner in which the experimental individuals are selected. To the extent that a simple structure and its associated primary factors are invariant under changes in selective conditions for the experimental group, we can have confidence that the primary factors represent identifiable processes whose nature transcends the circumstances under which the experimental subjects happen to be selected. It does not follow that these identifiable and interrelated processes are therefore unique in the mathematical sense of being the only set of parameters that can be used for describing the dynamic system that produces the observed individual differences. The invariance of simple structure under changes in selective conditions does imply, however, that the processes so identified are likely to be fruitful landmarks in a logical description of the system.

By *univariate selection* is meant the selection of a group of experimental subjects so that their standard deviation in a selection variable shall be different from unity, which is the dispersion in a set of normalized scores for an initial group of subjects. The criterion or selection variable may be defined by one of the tests in the battery or it may be a linear combination of several tests or factors of the initial analysis. If the selection criterion is correlated with any of the tests in the battery, then there will be an effect on the correlations between the tests in the selected group of subjects. The expected correlations in the new group can be computed by Godfrey Thomson's equation (18). The new communalities can be determined in an analogous manner.

When the new group of subjects is assembled according to dispersion in a selection variable L so that its standard deviation s_l satisfies the inequality $0 \neq s_l \neq 1$, then the new group is assembled by *partial selection* in the sense that variance in the selection variable is not entirely excluded but is different from that of the initial group of subjects. When the new group is assembled so that its dispersion in the selection variable vanishes entirely, then its standard deviation $s_l = 0$ and the new group is then described by *complete or total selection* on the selection variable.

The rank of the reduced correlation matrix for the new group is the same as for the initial group of subjects except when the selection is complete on the selection variable, in which case the dimensionality is reduced by one. We then have first-order partial correlation coefficients with the selection variable constant and the rank r is then reduced to $(r - 1)$.

In partial selection when the rank of the reduced correlation matrix is invariant, the simple structure is also invariant but the correlations between the primary parameters are altered. The interpretation of the primary factors remains the same in both groups.

In total selection when the rank of the reduced correlation matrix is reduced by one, the simple structure may be distorted depending on the relation of the total selection to the primary factors. The interpretation of the primary factors affected by the total selection may then be obscured.

When the new group is selected so as to satisfy conditions on two independent selection variables, the variances of the selection variables are affected and also their covariances. Such a situation is called *multivariate selection.* Multivariate selection can be described in terms of successive univariate selections. The expected correlations after multivariate selection can be computed by matrix formulas derived by Aitken. The new correlations can also be determined by the univariate formula of Godfrey Thomson applied to represent successive univariate selections.

With partial multivariate selection of a new group of subjects the rank of the reduced correlation matrix is augmented by one or more incidental factors. The number of incidental selection factors is determined by the number of independent variables that participate in the multivariate selection. The simple structure remains invariant so that the primary factors can be identified with the same interpretation as for the initial group of subjects. The additional factors may show appreciable variance but they will not be invariant for differently selected groups since they are determined by the conditions of selection of each group. If the attempt is made to interpret the incidental selection factors as basic parameters, then the interpretation will fail to be sustained in subsequent factorial studies of the same domain with differently selected subjects. The primary factors should be identified in the differently selected groups.

With total multivariate selection the rank of the correlation matrix is reduced by one or more factors depending on the number of successive variables on which the selection is complete, but the rank is also augmented by incidental selection factors depending on the number of selection variables that participate in the selection. With total selection involving the primary factors, the simple structure can be so distorted that the primaries may not be identified. For this reason it is well to allow as much variation as possible among the subjects in the domain which is to be investigated, thus practically eliminating the possibility of total selection within the parameters that are to be sought for. In factorial investigation it is not of any consequence whether any of the groups of experimental subjects are representative of a general population. The important consideration is that the experimental subjects should vary among themselves as much as possible within the domain that is being investigated.

The analysis of these various cases of selection is very encouraging in that a simple structure has been shown to be invariant under widely different selective conditions. Hence, the scientific interpretation of the primary factors as meaningful parameters can be expected to transcend the widely different selective conditions of the objects that are measured and factorially analyzed. This encouraging finding leads to a recommendation for the factorial study of any domain. When a simple structure has been found for a test battery that has been given to an experimental population and when a plausible interpretation of the primary factors has been found, these should be regarded as hypotheses to be verified by giving the same test battery to new experimental populations that should be selected in different ways. If the primary factors are in the nature of basic parameters that are not merely reflections of the experimental conditions or the particular selective conditions, then these factors should be invariant under widely different selective conditions and their interpretation should be the same for the several experimental groups. New test batteries should be constructed with prediction as to factorial composition of the new tests and these should be tried on differently selected subjects in order to determine whether the interpretation of each factor as a meaningful parameter can be sustained.

This is another situation in which the factorial methods depart from the conventions of statistical analysis. It is customary in statistical reasoning to think of a general population from which we merely draw samples and the worry is then whether the sample is really representative of the universe. In factor analysis, the principal concern is to discover an underlying order to be described in terms of meaningful parameters which should represent scientific concepts. The validity of a primary factor is determined by its fruitfulness as a scientific concept. It is inadequate if it serves merely as a regression coefficient. When factor analysis has completed its job of charting a new domain, then it may be of practical importance to establish norms of performance for any specified general population. Then the conventional statistical reasoning is again applicable.

TABLE 1

	A	*B*	Unique factors				
1	x	x	x				
2	x	x		x			
3	x	x			x		
4	x	x				x	
5	x	x					x
1*a*	x	x	x				

TABLE 2

	A	*B*	1	2	3	4	5
1	x	x	x				
2	x	x		x			
3	x	x			x		
4	x	x				x	
5	x	x					x
(1 + 2)	x	x	x	x			

TABLE 3
Given Correlation Matrix R_1

	1	2	3	4	5	6	7	8	9	10
1	1.000	.000	.000	.350	.606	.350	.606	.000	.000	.404
2	.000	1.000	.000	.606	.350	.000	.000	.350	.606	.404
3	.000	.000	1.000	.000	.000	.606	.350	.606	.350	.404
4	.350	.606	.000	1.000	.606	.175	.303	.303	.525	.552
5	.606	.350	.000	.606	1.000	.303	.525	.175	.303	.552
6	.350	.000	.606	.175	.303	1.000	.606	.525	.303	.552
7	.606	.000	.350	.303	.525	.606	1.000	.303	.175	.552
8	.000	.350	.606	.303	.175	.525	.303	1.000	.606	.552
9	.000	.606	.350	.525	.303	.303	.175	.606	1.000	.552
10	.404	.404	.404	.552	.552	.552	.552	.552	.552	1.000

TABLE 4
Given Factor Matrix F_0

	I	II	III
1	.265	—.270	.546
2	.542	.632	.047
3	.542	—.424	—.473
4	.620	.448	.313
5	.566	.116	.555
6	.620	—.519	—.165
7	.566	—.514	.257
8	.740	—.051	—.386
9	.740	.336	—.195
10	.815	—.024	.047

TABLE 5
Computation of s_j

	r_{jl}	qr_{jl}	$q^2r^2_{jl}$	s^2_j	s_j	m_j
1	1.000	.800	.640	.360	.600	1.667
2	.000	.000	.000	1.000	1.000	1.000
3	.000	.000	.000	1.000	1.000	1.000
4	.350	.280	.078	.922	.960	1.042
5	.606	.485	.235	.765	.875	1.144
6	.350	.280	.078	.922	.960	1.042
7	.606	.485	235	.765	.875	1.144
8	.000	.000	.000	1.000	1.000	1.000
9	.000	.000	.000	1.000	1.000	1.000
10	.404	.323	.105	.896	.946	1.057

$p = .60$
$p^2 = .36$
$q^2 = .64$
$q = .80$
$t = .40$
$t^2 = .16$
$s^2_j = 1 - q^2r^2_{jl}$
$q^2 = 1 - p^2$
$t = 1 - p$
$u_j = qr_{jl}$
$m_j = 1/s_j$

TABLE 6
Covariance Matrix $C_s = R_1 - UU'$ or $c_s = r_{jk} - u_j u_k$.

	1	2	3	4	5	6	7	8	9	10
1	.360	.000	.000	.126	.218	.126	.218	.000	.000	.146
2	.000	1.000	.000	.606	.350	.000	.000	.350	.606	.404
3	.000	.000	1.000	.000	.000	.606	.350	.606	.350	.404
4	.126	.606	.000	.922	.470	.097	.167	.303	.525	.462
5	.218	.350	.000	.470	.765	.167	.290	.175	.303	.395
6	.126	.000	.606	.097	.167	.922	.470	.525	.303	.462
7	.218	.000	.350	.167	.290	.470	.765	.303	.175	.395
8	.000	.350	.606	.303	.175	.525	.303	1.000	.606	.552
9	.000	.606	.350	.525	.303	.303	.175	.606	1.000	.552
10	.146	.404	.404	.462	.395	.462	.395	.552	.552	.896

TABLE 7
New Correlation Matrix $R_s = M\ C_s\ M$

	1	2	3	4	5	6	7	8	9	10
1	1.000	.000	.000	.219	.416	.219	.416	.000	.000	.256
2	.000	1.000	.000	.632	.400	.000	.000	.350	.606	.427
3	.000	.000	1.000	.000	.000	.632	.400	.606	.350	.427
4	.219	.632	.000	1.000	.560	.105	.199	.316	.547	.508
5	.416	.400	.000	.560	1.000	.199	.379	.200	.347	.478
6	.219	.000	.632	.105	.199	1.000	.560	.547	.316	.508
7	.416	.000	.400	.199	.379	.560	1.000	.347	.200	.478
8	.000	.350	.606	.316	.200	.547	.347	1.000	.606	.583
9	.000	.606	.350	.547	.347	.316	.200	.606	1.000	.583
10	.256	.427	.427	.508	.478	.508	.478	.583	.583	1.000

TABLE 8

New Communalities: $_{s}h^{2}_{j} = \dfrac{h^{2}_{j} - q^{2}\, r^{2}_{jl}}{1 - q^{2}\, r^{2}_{jl}}$

	h^2_j	r_{jl}	r^2_{jl}	$q^2r^2_{jl}$	Num.	Denom.	$_sh^2_j$
1	700	.837	.700	.448	.252	.552	.457
2	.700	.000	.000	.000	.700	1.000	.700
3	.700	.000	.000	.000	.700	1.000	.700
4	.700	.126	.016	.010	.690	.990	.697
5	.700	.218	.048	.031	.670	.970	.691
6	.700	.126	.016	.010	.690	.990	.697
7	.700	.218	.048	.031	.670	.970	.691
8	.700	.000	.000	.000	.700	1.000	.700
9	.700	.000	.000	.000	.700	1.000	.700
10	.700	.146	.021	.014	.686	.986	.696

TABLE 9
New Factor Matrix F_s

	I	II	III
1	.334	—.340	.688
2	.544	.634	.048
3	.543	—.425	—.474
4	.638	.379	.385
5	.561	.023	.619
6	.637	—.538	—.067
7	.561	—.507	.359
8	.742	—.051	—.387
9	.742	.337	—.196
10	.820	—.075	.151

TABLE 10
Given
Factor
Matrix

	I
1	.9
2	.8
3	.7
4	.6
5	.5
6	.4

TABLE 11
Given Correlation Matrix

	L	*M*	1	2	3	4	5	6
L	1.0000	.0000	.3741	—.3741	.0935	.0802	.0668	.0534
M	.0000	1.0000	.9274	.9274	.6416	.5500	.4583	.3667
1	.3741	.9274	1.00	.72	.63	.54	.45	.36
2	—.3741	.9274	.72	1.00	.56	.48	.40	.32
3	.0935	.6416	.63	.56	1.00	.42	.35	.28
4	.0802	.5500	.54	.48	.42	1.00	.30	.24
5	.0668	.4583	.45	.40	.35	.30	1.00	.20
6	.0534	.3667	.36	.32	.28	.24	.20	1.00

TABLE 12
Covariance Matrix After Selection on *M*

	L	*M*	1	2	3	4	5	6
L	1.0000	.0000	.3741	—.3741	.0935	.0802	.0668	.0534
M	.0000	.3835	.3557	.3557	.2460	.2109	.1757	.1406
1	.3741	.3557	.4699	.1899	.2632	.2255	.1880	.1504
2	—.3741	.3557	.1899	.4699	.1932	.1655	.1380	.1104
3	.0935	.2460	.2632	.1932	.7462	.2024	.1687	.1350
4	.0802	.2109	.2255	.1655	.2024	.8135	.1446	.1157
5	.0668	.1757	.1880	.1380	.1687	.1446	.8705	.0964
6	.0534	.1406	.1504	.1104	.1350	.1157	.0964	.9171

$$p = \frac{.5744}{.9274} = .6193$$

$$p^2 = .3835$$

$$q^2 = .6165$$

$$q = .7852$$

TABLE 13
Covariance Matrix After Selection on *L*

	L	*M*	1	2	3	4	5	6
L	.2145	.0000	.9802	—.0802	.0200	.0172	.0143	.0115
M	.0000	.3835	.3557	.3557	.2460	.2109	.1757	.1406
1	.0802	.3557	.3599	.2999	.2357	.2019	.1684	.1347
2	—.0802	.3557	.2999	.3599	.2207	.1891	.1576	.1261
3	.0200	.2460	.2357	.2207	.7393	.1965	.1638	.1311
4	.0172	.2109	.2019	.1891	.1965	.8084	.1404	.1123
5	.0143	.1757	.1684	.1576	.1638	.1404	.8670	.0936
6	.0115	.1406	.1347	.1261	.1311	.1123	.0936	.9149

$$p = \frac{.1732}{.3741} = .4630$$

$$p^2 = .2144$$

$$q^2 = .7856$$

$$q = .8863$$

TABLE 14
New Correlation Matrix

	1	2	3	4	5	6
1	.87	.83	.46	.38	.31	.24
2	.83	.78	.43	.35	.29	.22
3	.46	.43	.31	.26	.21	.16
4	.38	.35	.26	.21	.17	.13
5	.30	.28	.21	.17	.14	.11
6	.24	.22	.16	.13	.11	.08

TABLE 15
New Factor Matrix

	I	II
1	.82	.45
2	.76	.45
3	.56	0
4	.46	0
5	.37	0
6	.29	0

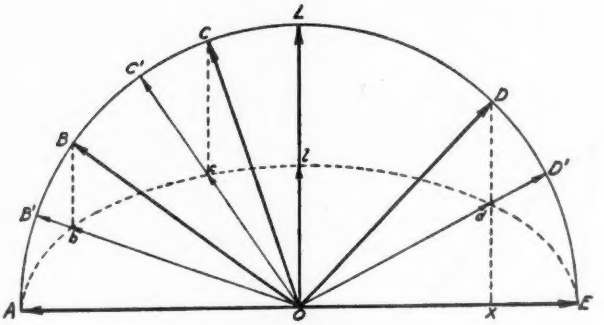


FIGURE 1

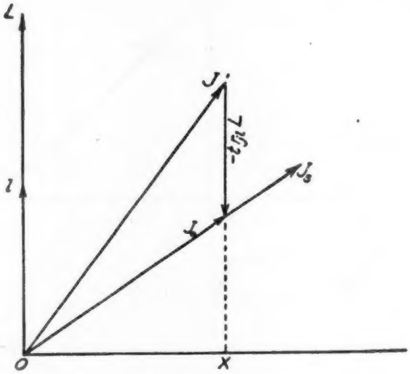


FIGURE 2

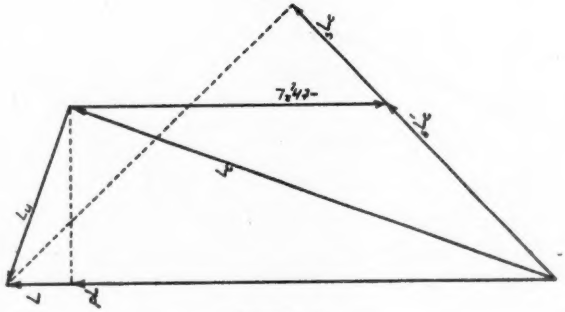


FIGURE 3

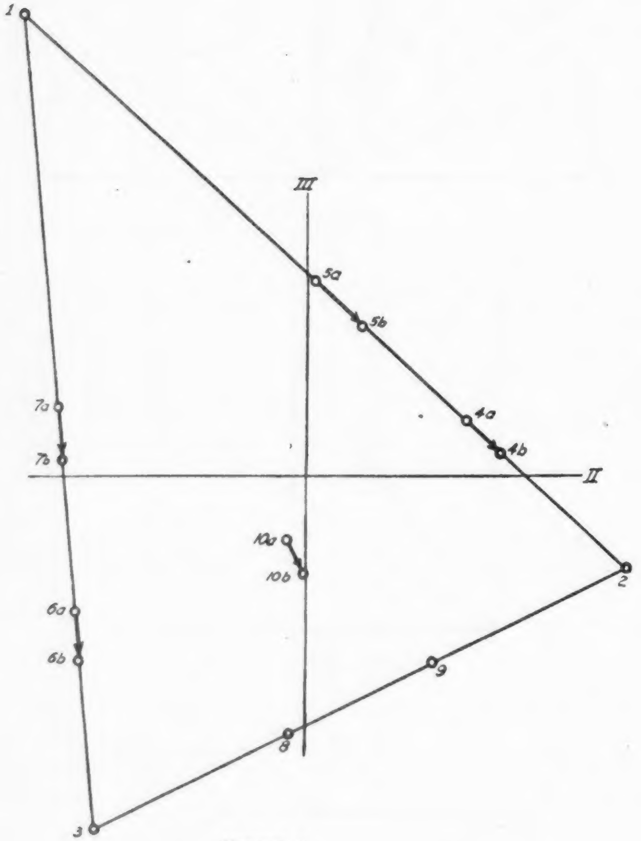


FIGURE 4

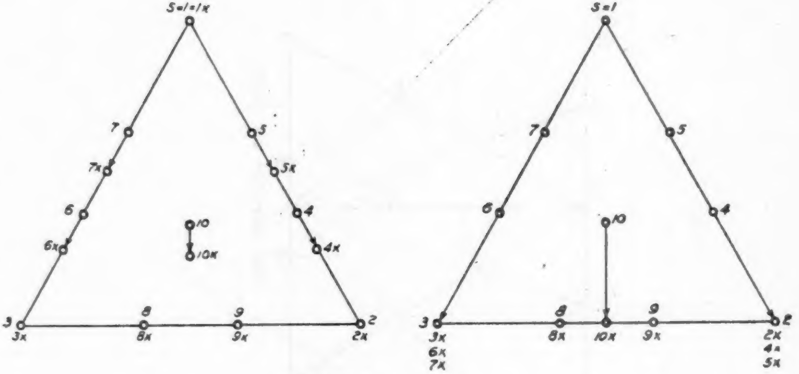
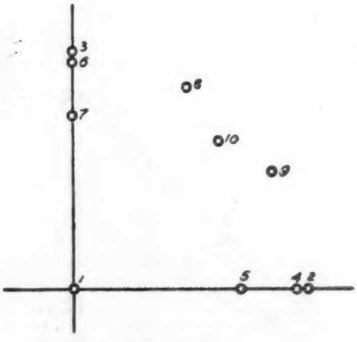


FIGURE 5



Total Selection on P_1

FIGURE 6

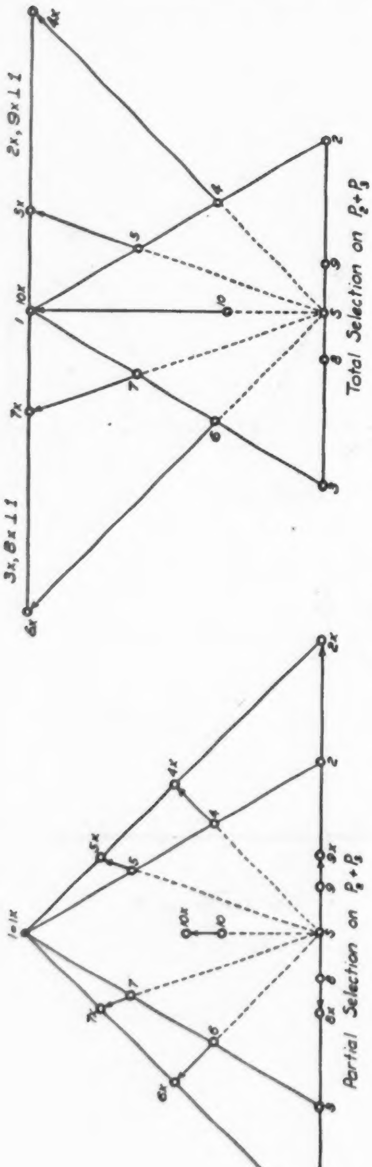
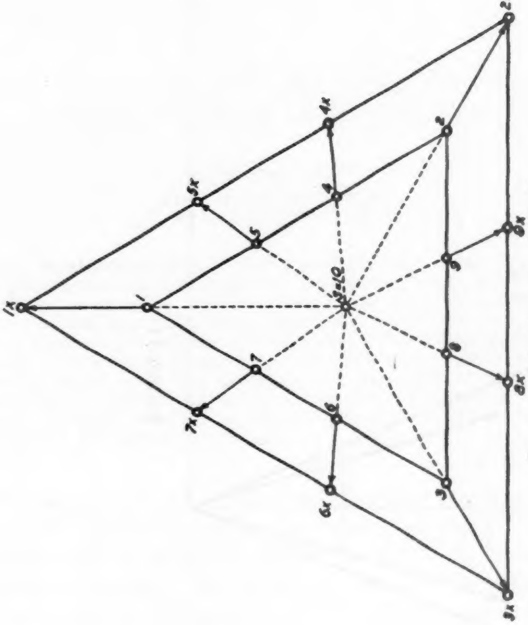
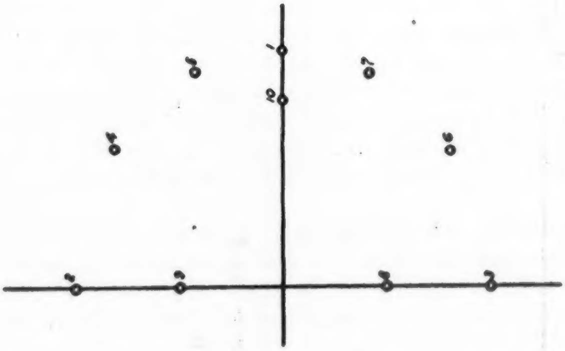

4x
2x, 9x ⊥ 1
5x
1 10x
7x
3x, 8x ⊥ 1
6x
4
5
10
7
6
2
9
3
8
3
Total Selection on P_2+P_3
1=1x
5x
4x
5
4
7x
7
10x
10
6x
6
2x
2
9x
9
3
8
8x
3
3x
Partial Selection on P_2+P_3

FIGURE 7



Partial Selection on $P_1+P_2+P_3$

FIGURE 9



Total Selection on P_1+P_2

FIGURE 8

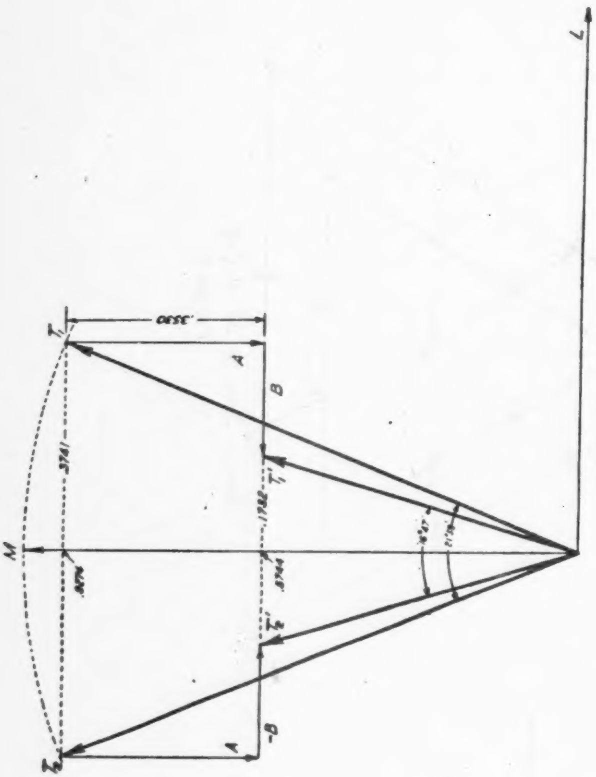
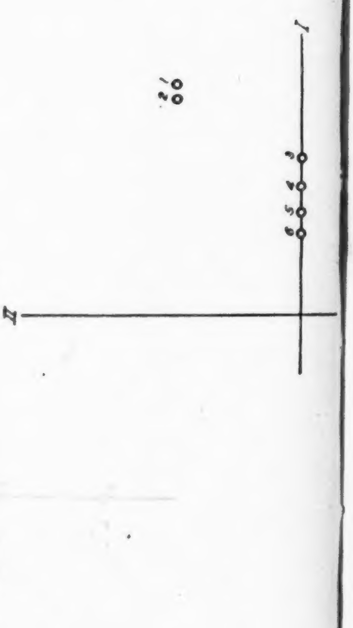
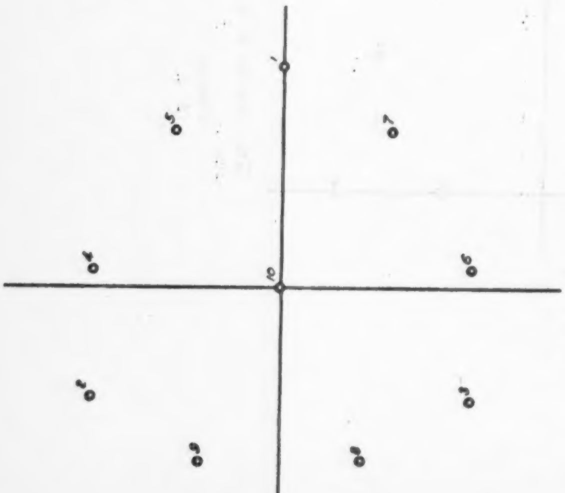


FIGURE 11





Total Selection on $P_1+P_2+P_3$

FIGURE 10

PSYCHOMETRIKA—VOL. 10, NO. 3
SEPTEMBER, 1945

TESTING LINEAR HYPOTHESES ILLUSTRATED BY A SIMPLE EXAMPLE IN CORRELATION

CYRIL J. HOYT
STATE TEACHERS COLLEGE
MANKATO, MINNESOTA

The development of a criterion suitable for testing the significance of a correlation or regression coefficient is used as an illustration of the manner in which a research problem is bound to the selection of the particular data appropriate to collect and a fitting type of statistical analysis of the latter. The translation of the original inquiry into a problem of "testing linear hypotheses" is the means by which these two aspects of an investigation are held together. This presentation is offered as a plan which might be useful for some research workers in determining appropriate criteria for testing their particular hypotheses.

In 1936, Johnson and Neyman (2) discussed the manner in which several types of problems frequently arising in educational research could be attacked in a scientific manner by applying sound principles of experimental design. They pointed out that the particular data which the investigator would find useful and the appropriate manner of analysis of the same were two aspects of a study both of which were determined by a careful translation of the problem into the precise language of mathematics. In their discussion of this latter phase, these authors disclosed that the appropriate analysis, in many cases, would fall in the category of "testing linear hypotheses."* They applied this method of treatment to an important type of problem in the field of educational research and then indicated how the same general mode of approach could be employed in a number of other situations.

Specifically, these authors considered the manner of comparing two groups of students in some measurable character when there were known inequalities between the groups on two qualities presumably related to the former. Following the suggestion of the earlier paper, the writer has elsewhere (1) treated, by the method presented below, the more general problem of comparing any number of groups after making adjustments in the criterion for inequalities existing in any number of related qualities. In addition, he has illustrated how

* As Johnson and Neyman (2) state clearly, to Kolodziejczyk (3) must go the utmost credit for his fundamental work on this general problem.

the category of "testing linear hypotheses" includes such inquiries as the significance of any partial regression coefficient, any combination of them, or the variation among the values of these statistics found in several samples.

In order to clarify the procedure followed in the present paper, it is necessary to explain the meaning of hypothesis and the logic underlying their tests. Certainly, one of the most frequent as well as most important purposes for the analysis of data consists in enabling the research worker to infer certain characteristics of the population of which the body of observations may be considered a sample. The law governing the manner in which the individuals of the population are distributed is often so closely approximated by some mathematical formula that properties of this function are useful for further study of the population. With these facts in mind, it is now possible to define a statistical hypothesis. The latter term refers to any statement designating the functional form with or without specifying any or all of the parameters of this law of distribution. Tests of various types have been devised to aid the investigator in quantifying the degree of his confidence that a certain population is the one sampled.

Probably one of the most significant features of the approach which Kolodziejczyk (3) and later Johnson and Neyman (2) present is the convenient and precise manner in which the practical research problem is translated into a mathematical one which can, in turn, be treated by more or less routine methods of mathematics. It is true that other approaches to the analysis of variance and covariance yield the same results in problems of "testing significance" as does the one employing tests of linear hypotheses, but the *choice* of the appropriate statistical tool for the research problem at hand is usually not as precise as that made through the latter approach. Often this choice of the appropriate statistical tool depends upon the research worker's intuitional grasp of the problem and of the available statistical tools. As has been true in the past, intuition will probably render inestimable aid to the research workers of the future. However, a more methodical and routinized translation of a research problem into a mathematical one possesses important advantages in the solution of really new problems, even for the expert, and certainly has vital advantages for those whose intuition may not be unfailing.

A statistical hypothesis was designated as "linear" if the specified population mean was that sort of a function of one or more parameters and if the deviations from this mean were assumed to be normally distributed (in the population) with a common variance. Tests based upon the principle of likelihood were developed **(1, 2, 3, 4)** for several hypotheses of this type, including those mentioned in the second paragraph above. The method of determining the appropriate criterion for one of these tests is illustrated by the simple example below.

The investigator may advance the tentative theory that, among fifth-grade children, there is a linear relation between scores on a certain reading test (y) and those on an examination on the accuracy of arithmetical computation (x). If the scores on these tests are considered to have been reduced to deviations from their respective means, the equation for translating the above problem into algebraic symbols is:

$$y = bx + z\,, \tag{1}$$

where b is a constant to be determined and z denotes that portion of y not linearly associated with x. It is these quantities (z) which the investigator is willing to assume are distributed normally in the population from which the observed sample was drawn. Then the probability of the occurrence of the observed sample subject to the assumption that b and σ have some specified values is

$$p\{(z_1, \cdots, z_n)/H\} = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e - \frac{\sum z^2}{2\sigma^2}, \tag{2}$$

where n denotes the number in the observed sample, $\sum$ indicates the sum for all n observations and H denotes the whole class of simple hypotheses any one of which we are willing to accept alternative to H'', the one we wish to test. The latter in its null form would be stated thus: There is no significant linear association between the scores on these particular reading and arithmetic tests in the population from which the observed sample was drawn in a random manner. In terms of our symbols, this hypothesis would be stated as

$$b = 0\,. \tag{3}$$

The logarithm of p, Equation (2), is:

$$\log p = -n \log \sigma - n \log \sqrt{2\pi} - \frac{\sum (y - bx)^2}{2\sigma^2}. \tag{4}$$

In order to determine the values of σ and b which maximize (2), we differentiate (4) partially with respect to σ and b, where we consider the latter as continuous variables and the sums of the x's and the y's as fixed.

$$\frac{\partial \log p}{\partial \sigma} = -\frac{n}{\sigma} + \frac{\sum (y - bx)^2}{\sigma^3}; \tag{5}$$

$$\frac{\partial \log p}{\partial b} = -\frac{1}{\sigma^2}\Sigma(y - bx)x. \tag{6}$$

When these are each set equal to zero, we obtain the following system of equations:

$$\Sigma\, xy - b\,\Sigma x^2 = 0 \quad \text{or} \quad b = \frac{\Sigma xy}{\Sigma x^2}; \tag{7}$$

$$n\,\sigma^2 = \Sigma(y - bx)^2 \quad \text{or} \quad n\,\sigma^2 = \Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2}. \tag{8}$$

This quantity in the right member is recognized as the sum of the squares of the residual deviations from the regression line. Denote this by nS^2_a.

In the next step in the process of obtaining the appropriate criterion for testing the null hypothesis stated above, we must consider the modification in (2) associated with the presence of the value zero for b as proposed by H''.

$$p''\{(z_1, \cdots, z_n)/H''\} = \left(\frac{1}{\sigma''\sqrt{2\pi}}\right)^n e - \frac{\Sigma y^2}{2\sigma''^2}. \tag{9}$$

Here the logarithm of p'' is:

$$\log p'' = -n \log \sigma'' - n \log \sqrt{2\pi} - \frac{\Sigma y^2}{2\sigma''^2}. \tag{10}$$

The only parameter involved in (10) is σ'', so we determine the minimum value of (10) by differentiating with respect to σ'', setting the derivative equal to zero and solving the resulting equation.

$$\frac{\partial \log p''}{\partial \sigma''} = -\frac{n}{\sigma''} + \frac{\Sigma y^2}{\sigma''^3}, \tag{11}$$

$$n\,\sigma''^2 = \Sigma y^2\,. \tag{12}$$

Denote the quantity in the right member of (12) by nS^2_r. Hence from (12) we have $\sigma''^2 = S^2_r$ and similarly, from (8), we have $\sigma^2 = S^2_a$. Substituting these values in (2) and in (9), we have the maximum values of these two functions.

$$p(\max - H) = \left(\frac{1}{S_a\sqrt{2\pi}}\right)^n e - \frac{n}{2}; \tag{13}$$

$$p''(\max - H'')\left(\frac{1}{S_r\sqrt{2\pi}}\right)^n e - \frac{n}{2}. \tag{14}$$

The ratio

$$\lambda = \frac{p''(\max - H'')}{p(\max - H)} \tag{15}$$

was defined by Neyman and Pearson (4) as the likelihood of H'' when tested against the whole set of alternatives H. Thus the likelihood of any hypothesis such as H'' is the ratio of the probability that the observed sample occurred subject to the conditions specified by H'', to its probability subject to the conditions specified by that member of the set of alternatives which renders this probability greater than does any other member of the set. In this case the ratio of (13) to (14) is

$$\lambda = \left(\frac{S_a}{S_r} \right)^n \quad \text{or} \quad \left(\frac{S^2{}_a}{S^2{}_r} \right)^{n/2}. \tag{16}$$

In practice, it has become customary to use U as a criterion instead of λ, where

$$U = \lambda^{2/n} \quad \text{or} \quad U = \frac{nS^2{}_a}{nS^2{}_r}. \tag{17}$$

Among others, Kolodziejczyk (3) has shown that under the assumption that random samples are chosen from the same normal population, this ratio U is distributed as the incomplete beta function. Since the integral of this function has been tabled (5), it is possible to determine the value of U corresponding to the particular probability level (ε) chosen as the risk we are willing to take of rejecting a hypothesis that correctly specifies the parameters of the population sampled. Johnson and Neyman (2) have published an extension of Pearson's table (5) which includes the values required for (n—s) greater than 100, where n represents the number in the observed sample and s the number of independent parameters of which the population mean is assumed to be a linear function.

In the case of this particular example, the investigator would calculate the value of (17) by computing the ratio of

$$\left(\Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2} \right) \text{ to } \Sigma y^2.$$

Then he would enter Pearson's table (5) if n were less than 102 or Johnson and Neyman's (2) if n were 102 or greater. If he found the observed value of (17) was greater than the one given in the appropriate table corresponding to the probability level (ε), he would accept the null hypothesis. It is clear that (17) would be equal to unity if the correlation between x and y were zero and, likewise, (17) would decrease with an increased numerical value of that coefficient of correlation.

The foregoing may seem to be a rather complicated approach to the test of the significance of the ordinary correlation coefficient. However, this is presented as an illustration of the systematic manner in which the appropriate criterion can be developed. Furthermore, this illustrates how the research worker who approaches his problem from the point of view of testing linear hypotheses is led to the collection of appropriate data and a fitting choice of the mode of statistical analysis. The author cannot claim these ideas as original but offers this illustration to the reader who may not be in a position to read the original papers.

REFERENCES

1. Hoyt, Cyril J. Tests of certain linear hypotheses and their application to educational problems in elementary college physics. Unpublished Ph.D. thesis, Graduate School, University of Minnesota, Minneapolis, (1944).
2. Johnson, Palmer O. and Neyman, J. Tests of certain linear hypotheses and their application to some educational problems. *Statistical Research Memoirs*, 1936, 1, 57-93.
3. Kolodziejczyk, S. On an important class of statistical hypotheses. *Biometrika*, 1935, 27, 161-190.
4. Neyman, J. and Pearson, E. S. On the use and interpretation of certain test criteria for the purpose of statistical inference. *Biometrika*, 1928, XXA, 175-240 and 263-294.
5. Pearson, K. Tables of the incomplete beta function. Biometrika Office, University College, London.

PSYCHOMETRIKA—VOL. 10, NO. 3
SEPTEMBER, 1945

APPROXIMATE METHODS IN CALCULATING DISCRIMINANT FUNCTIONS

GEOFFREY BEALL
INSTITUTE OF PAPER CHEMISTRY, APPLETON, WISCONSIN*

Approximate methods of solving for discriminant functions have been tried on three sets of data. The principal illustration is the problem of finding a weighted sum of scores, on four psychological tests, so that men and women may be distinguished most clearly. The work starts from the complete solution, due to R. A. Fisher, where it is necessary to solve as many simultaneous equations, dependent on the standard deviations of the tests and their mutual correlations, as there are tests. It is proposed, by way of numerical simplification, that a set of equations be substituted where some one quantity replaces all the correlations. A solution is obtained where the weights to be assigned the tests are very simply expressed in terms of differences between the mean values of tests, the standard deviations of tests, and the said quantity. The difficulty remains of finding an estimate of the arbitrary constant that will give good discrimination. If an optimal solution is made a result is obtained which, in the three sets of data considered, is almost indistinguishable from that yielded by the complete solution. The calculation of this optimal common quantity is, however, itself so considerable that another estimate, previously suggested by R. W. B. Jackson, appears more profitable. This estimate is derived simply from the variability between the total scores for each subject and the variability of each test. Using this estimate, the discriminant functions can be rapidly calculated; the results compare very favorably, in the case of the data considered, with those from the complete solution.

1. The Problem of Discriminating Between Two Groups

The problem of discrimination, i.e., of so combining various considerations on a given object (or subject) that objects belonging to one group may be characterized to distinguish them most clearly from those of an alternate group, has been treated, in a general way, by Fisher (1937). Fisher illustrated his method with the discrimination of two species of *Iris* on the basis of sepal and petal size. The statistical problem is the same when it is required to discriminate between, say, normal and psychotic men on the basis of various psychological tests. The present work is concerned with approximate methods, on the line suggested by Jackson (1943), which may be justified by their facilitating practice.

The data used to establish the method of discrimination must consist of an observation on the $ith (i = 1, \cdots, m)$ consideration of

* The present work was done while the writer was employed by the Ontario Department of Health.

the jth $(j = 1, \cdots, n_1 + n_2)$ object, as a score, x_{ij} when the objects have been assigned to Group I or Group II, respectively, for

$$\begin{aligned} 1 &\leqslant j \leqslant n_1 \\ n_1 + 1 &\leqslant j \leqslant n_1 + n_2 . \end{aligned} \tag{1}$$

Fisher (1937) has solved the problem of discrimination by supposing that there is required a linear function,

$$y_j = \sum_{i=1}^{m} a_i x_{ij} , \tag{2}$$

subsequently termed discriminant, where the values a_i are to be chosen in such proportion that D^2/S shall be maximal where, for the discriminant values, D^2 depends on the variability between groups and S on the variability within groups. Specifically, set

$$D = y_1 - y_2 = \sum_{i=1}^{m} a_i (x_{i1} - x_{i2}) , \tag{3}$$

where, for Group I ,

$$y_1 = \sum_{j=1}^{n_1} y_j / n_1 = \sum_{i=1}^{m} a_i x_{i1} \tag{4}$$

is the mean discriminant value for Group I and

$$x_{i1} = \sum_{j=1}^{n_1} x_{ij} / n_1 \tag{5}$$

is the mean value of the ith consideration for Group I and values y_2 and x_{i2} are similarly defined for Group II. Secondly,

$$S/(n_1 + n_2) = \sum_{i=1}^{m} a^2{}_i s^2{}_i + \sum_{i=1}^{m} \sum_{i'=1}^{m} a_i a_{i'} s_i s_{i'} r_{ii'} , \tag{6}$$

for $i' \neq i$, where

$$s^2{}_i = \frac{1}{n_1 + n_2} \left\{ \sum_{j=1}^{n_1} (x_{ij} - x_{i1})^2 + \sum_{j=n_1+1}^{n_1+n_2} (x_{ij} - x_{i2})^2 \right\} \tag{7}$$

is the standard deviation within groups for the ith consideration and where

$$\begin{aligned} r_{ii'} = \Bigg\{ & \sum_{j=1}^{n_1} (x_{ij} - x_{i1}) (x_{i'j} - x_{i'1}) \\ & + \sum_{j=n_1+1}^{n_1+n_2} (x_{ij} - x_{i2}) (x_{i'j} - x_{i'2}) \Bigg\} / (n_1 + n_2) s_i s_{i'} , \end{aligned} \tag{8}$$

is the corresponding correlation coefficient within groups. The required minimization is given by the proportional solution, for the quantities a_i, of the m equations,

$$a_i\, s_i + \sum_{i'=1.}^{m} a_{i'}\, r_{ii'}\, s_{i'} = t_i\,, \tag{9}$$

where

$$t_i = (x_{i1} - x_{i2})/s_i\,. \tag{10}$$

The work of setting up the numerical equations involved in (9) may be very heavy, since first $m(m-1)/2$ sums of squares or of cross-products must be found and then m equations must be solved. Indeed the procedure would probably become impracticable for m greater than about 8.

2. *Approximate Solutions*

To avoid the work required by a complete solution of such a set, for m great, of equations as (9), Jackson (1943) has suggested on empirical grounds an approximate solution based on the assumption that for all i and i'

$$r_{ii'} = r\,, \tag{11}$$

so that from (6)

$$S/(n_1 + n_2) = \sum_{i=1}^{m} a_i^2\, s_i^2 + r \sum_{i=1}^{m}\ \sum_{i'=1}^{m} a_i\, a_{i'}\, s_i\, s_{i'}. \tag{12}$$

If (12) be used instead of (6) and the ratio be minimized with respect to the coefficients a_i, we obtain, for any given value of r, the m equations,

$$a_i\, s_i + r \sum_{i'=1}^{m} a_{i'}\, s_{i'} = t_i\,, \tag{13}$$

in place of (9). The most convenient proportional solution of (13) is

$$a_i = \{(1-r)\,t_i + mrt'_i\}/s_i\,, \tag{14}$$

where

$$t'_i = t_i - t \tag{15}$$

and

$$t = \frac{1}{m} \sum_{i=1}^{m} t_i \tag{16}$$

is the mean value of t_i.

Equation (14) makes the coefficients dependent on one variable, r, which must still be determined for any data. The best value of

r may be found by substituting from (14) in (3) and (6) for a_i and then maximizing D^2/S with respect to r, to get

$$r = \frac{mt^2C + mtD - tAB - AC}{-m^2tA + mt^2C - m(m-1)tD - tAB + (m-1)AC}, \tag{17}$$

where

$$\begin{aligned} A &= \sum_{i=1}^{m} t'^2_i \\ B &= \sum_{i=1}^{m} \sum_{i'=1}^{m} r_{ii'} \\ C &= \sum_{i=1}^{m} \sum_{i'=1}^{m} t'_i r_{ii'} \\ D &= \sum_{i=1}^{m} \sum_{i'=1}^{m} t'_i t'_{i'} r_{ii'} . \end{aligned} \tag{18}$$

Obviously, (17) cannot be used in practice, for which a useful relation is suggested by the fact that insofar as (12) can be identified with (6),

$$r = \sum_{i=1}^{m} \sum_{i'=1}^{m} \left\{ a_i a_{i'} s_i s_{i'} r_{ii'} \right\} \Big/ \left\{ \sum_{i=1}^{m} \sum_{i'=1}^{m} a_i a_{i'} s_i s_{i'} \right\}, \tag{19}$$

where the coefficients, a_i, are not limited. It is convenient to make

$$a_i = 1 , \tag{20}$$

for all i, when from (19),

$$\begin{aligned} r &= \left\{ \sum_{i=1}^{m} \sum_{i'=1}^{m} r_{ii'} s_i s_{i'} \right\} \Big/ \left\{ \sum_{i=1}^{m} \sum_{i'=1}^{m} s_i s_{i'} \right\} \\ &= \left\{ s_0^2 - \sum_{i=1}^{m} s_i^2 \right\} \Big/ \left\{ s^2 - \sum_{i=1}^{m} s_i^2 \right\} , \end{aligned} \tag{21}$$

where

$$s_0^2 = \frac{1}{n_1 + n_2} \left\{ \sum_{j=1}^{n_1} (x_j - x_1)^2 + \sum_{j=n_1+1}^{n_1+n_2} (x_j - x_2)^2 \right\}, \tag{22}$$

where

$$x_j = \sum_{i=1}^{m} x_{ij} , \tag{23}$$

$$x_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} x_j , \tag{24}$$

and x_2, similarly defined, is the squared standard deviation within the groups for the total score of each object (or subject) on all considerations, and where

$$s = \sum_{i=1}^{m} s_i . \tag{25}$$

It will be seen that in (21) r is the mean of the quantities $r_{ii'}$, weighted by standard deviations.

The estimate of (21) is very similar to one previously proposed by Jackson (1943), i.e., his r'', which we shall write

$$r = \left\{ z_0^2 - \sum_{i=1}^{m} z_i^2 \right\} \Big/ \left\{ z^2 - \sum_{i=1}^{m} z_i^2 \right\} , \tag{26}$$

where

$$z_i^2 = \frac{1}{n_1 + n_2} \sum_{j=1}^{n_1+n_2} (x_{ij} - x_i)^2 \tag{27}$$

is the squared standard deviation over all the material for the ith consideration (embracing both groups),

$$z_0^2 = \frac{1}{n_1 + n_2} \sum_{j=1}^{n_1+n_2} (x_j - x)^2 , \tag{28}$$

when

$$x = \frac{1}{n_1 + n_2} \sum_{j=1}^{n_1+n_2} x_j \tag{29}$$

is the squared standard deviation over all the material for the total score of each subject, and

$$z = \sum_{i=1}^{m} z_i . \tag{30}$$

It will be seen that (26) differs from (21) in that the variability is calculated over all the $n_1 + n_2$ objects, rather than within the groups of n_1 and n_2, respectively. Jackson's estimate of (26) is related by

$$r = \frac{\sum_{i=1}^{m} \sum_{i'=1}^{m} s_i s_{i'} (r_{ii'} + t_i t_{i'}/4)}{\sum_{i=1}^{m} \sum_{i'=1}^{m} s_i s_{i'} \sqrt{(1 + t_i^2/4)(1 + t_{i'}^2/4)}} \tag{31}$$

to the values $r_{ii'}$.

TABLE 1
The Relative Success of Various Sets of Discriminant Coefficients

	Complete Solution (9)	On basis of common r			Limiting cases		Using $a_i = 1$
Example		Max. r of (17)	r of (21)	r of (26)	t'_i/s_i of (32)	t_i/s_i of (33)	
The ratio of the sum of squares between groups to the sum within groups							
Fisher (1937)	26.34	26.28	25.97	26.03	22.15	17.46	3.85
Travers (1939)	1.72	1.72	1.64	1.68	1.64	1.62	.43
Present Data	1.57	1.57	1.54	1.55	1.18	.89	.45
The square root of the ratio of the sum of squares between groups to the total sum of squares, or the coefficient, R, of multiple correlation.							
Fisher (1937)	.98	.98	.98	.98	.98	.97	.89
Travers (1939)	.80	.79	.79	.79	.79	.79	.55
Present Data	.78	.78	.78	.78	.74	.69	.56

For three problems of discrimination the satisfaction to be obtained by determining variously the coefficients, a_i, was found with the results shown in Table 1. First there are data due to Fisher (1937) on the discrimination of two species of *Iris, setosa* and *versicolor,* by measurements on length and breadth of sepals and petals of 50 specimens of each species. Secondly, there are data due to Travers (1939), on the discrimination of engineers and air pilots by the scores of 20 men in each group on 6 tests involving understanding, co-ordination, etc. These data of Travers have already been reconsidered by Jackson (1943). Thirdly, there are some fresh data,* as shown in Table 2, on psychological tests given 32 men and 32 women. For each lot of data, discriminant coefficients were found from the complete solution of (9), from the best value of r, as given by (17) and from the estimate of r, as given by (21). Further the estimated r, of Jackson (1943) as in (26) was used throughout because he had already used it similarly for Travers' data. Also the relation

$$a_i = t'_i/s_i\,, \tag{32}$$

which is approached as r or m increases in (14), and

$$a_i = t_i/s_i\,, \tag{33}$$

which obtains for $r = 0$ in (14), were used. Finally, a_i was determined as in (20). For each set of coefficients, for each lot of data, there was found the ratio of the sum of squares between means to that within groups; originally, as in connection with (2), it was desired to make this ratio maximal. In the present cases where

$$n_1 = n_2 = n \tag{34}$$

the ratio tabled is $n\,D^2/2S$. In addition the values of

$$R = \{n\,D^2/(2S + n\,D^2)\}^{\frac{1}{2}}, \tag{35}$$

as used by Jackson (1943), are shown. From Table 1, it is apparent that using r of (17) the results were almost as good as from the complete solution, although the coefficients were not identical. Using r of (21) a very good result was obtained; but Jackson's r'' of (26) worked even better in all three cases considered. The very convenient coefficient t'_i/s_i, of (32), gave a result not far from that obtained by (21) in (14), as was anticipated, since often in (14) the second term in the denominator must be predominant, when, effectively, (32) obtains. The equally convenient coefficient t_i/s_i, of (33), gave a poorer result. The result from (20) is not of particular interest except to show how bad discrimination may be when, as in school-room prac-

* The present data were kindly made available by Dr. L. S. Penrose.

tice, the results of examinations are combined by simple addition.

The encouraging performance of a common estimate of r may be appreciated better when one notes that for the present three examples the correlations for the various considerations were as shown in Table 2. The second case (Travers 1939) is rather a stringent test of the utility of an assumption of a common r because, as will be seen in Table 2, the sign of the quantities $r_{ii'}$ varies. Travers notes that his fifth test was scored in a negative sense, i.e., "a low score indicates a good performance." In practice, where one knew such a condition to exist, one would, presumably, reverse the sign of such scores before considering a common value of r . If such a reversal is made in the present case, (21) gives $r = + .09$ and the ratio, as in Table 1, is but little affected.

TABLE 2
The Correlations of the Various Considerations

Fisher (1937)

Measurement	1	2	3
2	+.60		
3	+.64	+.38	
4	+.47	+.46	+.71

Travers (1939)

Test	1	2	3	4	5
2	+.25				
3	+.05	+.08			
4	+.07	—.04	+.03		
5	—.02	—.41	—.05	—.02	
6	+.38	+.14	—.19	+.07	—.02

New Data

Test	1	2	3
2	+.57		
3	+.39	+.39	
4	+.37	+.31	+.55

The estimates of a common r from the two relations (17) and (21) together with Jackson's suggestion, r'' of (26), are shown in Table 3.

TABLE 3
Various Estimates of a Common Value r

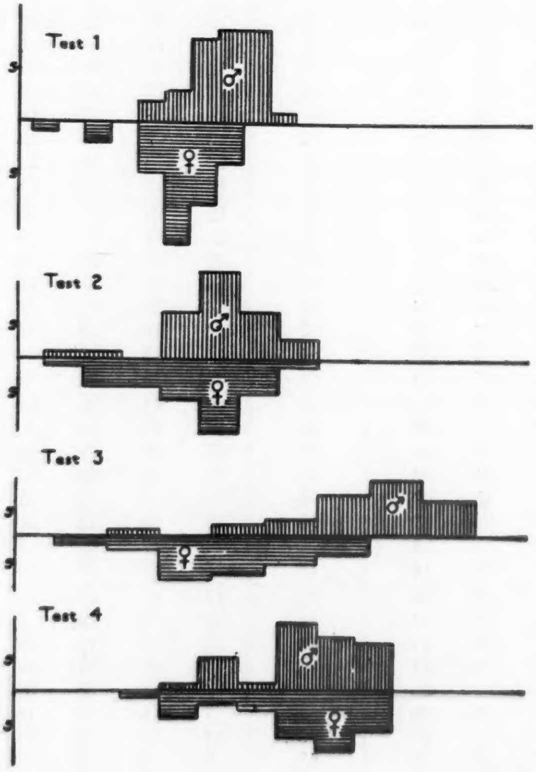
	r of (17)	r of (21)	r of (26)
Fisher (1937)	+.44	+.55	+.36
Travers (1939)	+.20	+.06	+.14
New Data	+.54	+.43	+.47

3. An Illustrative Example

The results, as summarized in Table 1, suggest that in practice an investigator might use as discriminant coefficients the values t'_i/s_i, as of (32), a common value of r, as from (26) in (14), or the complete solution as in (9), depending on circumstances. Below are illustrated in some detail the three alternative procedures to be followed and the results to be obtained by each procedure. The example chosen is the discrimination of men from women by four psychological tests, as first mentioned in connection with Table 1. Test 1 is on pictorial absurdities, 2 on paper form board, 3 on tool recognition, and 4 on vocabulary. The data on women are for 32 applicants for a profes-

FIGURE 1*

The Distribution of the Original Measurements on Four Tests, by Sex.



[*] The present illustrations are due to Mr. J. A. McClure of the Ontario Department of Health.

sional position requiring 10 or more years of successful schooling (the completion of second-year high school in Ontario, up to a University degree). Against these data were set results for men chosen with the same academic restrictions; from a very large group, 32 men were drawn randomly. The 4 tests were each scored according to the number of questions answered successfully. The correlations between the tests are shown in Table 2. The data are set out for reference in Table 4. The distribution of the original measurements by sex may be of some interest and is set out in Figure 1. It will be noted that the

TABLE 4

The Scores of 32 Men and 32 Women on Four Psychological Tests

Men					Women				
1	2	3	4		1	2	3	4	
15	17	24	14	70	13	14	12	21	60
17	15	32	26	90	14	12	14	26	66
15	14	29	23	81	12	19	21	21	73
13	12	10	16	51	12	13	10	16	51
20	17	26	28	91	11	20	16	16	63
15	21	26	21	83	12	9	14	18	53
15	13	26	22	76	10	13	18	24	65
13	5	22	22	62	10	8	13	23	54
14	7	30	17	68	12	20	19	23	74
17	15	30	27	89	11	10	11	27	59
17	17	26	20	80	12	18	25	25	80
17	20	28	24	89	14	18	13	26	71
15	15	29	24	83	14	10	25	28	77
18	19	32	28	97	13	16	8	14	51
18	18	31	27	94	14	8	13	25	60
15	14	26	21	76	13	16	23	28	80
18	17	33	26	94	16	21	26	26	89
10	14	19	17	60	14	17	14	14	59
18	21	30	29	98	16	16	15	23	70
18	21	34	26	99	13	16	23	24	76
13	17	30	24	84	2	6	16	21	45
16	16	16	16	64	14	16	22	26	78
11	15	25	23	74	17	17	22	28	84
16	13	26	16	71	16	13	16	14	59
16	13	23	21	73	15	14	20	26	75
18	18	34	24	94	12	10	12	9	43
16	15	28	27	86	14	17	24	23	78
15	16	29	24	84	13	15	18	20	66
18	19	32	23	92	11	16	18	28	73
18	16	33	23	90	7	7	19	18	51
17	20	21	21	79	12	15	7	28	62
19	19	30	28	96	6	5	6	13	30
511	509	870	728	2618	395	445	533	702	2075

main discrimination is by Test 3.

The discrimination to be undertaken would depend in large part on practical considerations. If the data were preliminary or the investigator were pressed, he might decide to use simply the values t'_i/s_i (multiplied by 10 in the present case) as discriminant coefficients and could anticipate reasonably good results. Then for the data of Table 4, he would need only make the calculations as shown in the first panel of Table 5. If the investigator could go to more pains, he might estimate, using (26) from the data of Table 4,

$$r = \frac{231.12669 - 104.57690}{(19.40928)^2 - 104.576090} = +.4650,$$

where

$$z_0^2 = 231.12669$$

is calculated from the totals over all 4 tests for individuals, i.e., from the 5th and 10th columns of Table 4;

$$\sum_{i=1}^{m} z_i^2 = 104.57690$$

is based on the preliminary calculation of standard deviations within the 1st and 6th columns, etc., of Table 4; and

$$z = 19.40928,$$

the mean standard deviation, is derived similarly. Using the estimate, $r = +.4650$, in (14), the calculations shown in the second panel of

TABLE 5
The Calculation of Discriminant Coefficients for Two Approximate Solutions
The coefficients $a_i = t'_i/s_i$ of Equation (32)

Test	t_i	t'_i	s_i	$a_i = t'_i/s_i$
1	1.3760	+.3672	2.6345	+1.3938
2	.5097	—.4991	3.9291	—1.2703
3	1.9748	+.9660	5.3328	+1.8114
4	.1747	—.8341	4.6501	—1.7937

The coefficients a_i on the common r of (27) in (14)

Test	$(1-r)t_i$	mrt'_i	$a_is_i =$ $(1-r)t_i + mrt'_i$	a_i
1	+ .7362	+ .6830	+1.4192	+5.3870
2	+ .2727	— .9283	— .6556	—1.6686
3	+1.0565	+1.7968	+2.8533	+5.3505
4	+ .0935	—1.5514	—1.4579	—3.1352

Table 5 are necessary. The values, a_i, again include a factor of 10. For both the sets of a_i, the ratio of sum of squares between and within groups is, of course, shown in Table 1. The discriminant distributions have been, further, plotted in Figure 2, to aid in appreciation of the discrimination obtained.

Finally, conditions might be such that the investigator would elect to make the complete solution of (9). It would then be necessary to find the values s_i and $r_{ii'}$ $s_{i'}$ from (7) and (8). The quantities $n^2 r_{ii'} s_i s_{i'}$ and $n^2 s_i^2$ (multiples by n of the sums of squares and cross-products about group means), as shown in Table 6, are most conveniently first found. Substituting from this table in (9), a proportional solution is as follows:

TABLE 6

Values of $n^2\, s_i^2$ and $n^2\, r_{ii'}\, s_i\, s_{i'}$ Required for the Complete Discriminant

Test	1	2	3	4
1	14,214			
2	11,998	31,534		
3	11,295	16,849	58,243	
4	9,326	11,618	27,738	44,284

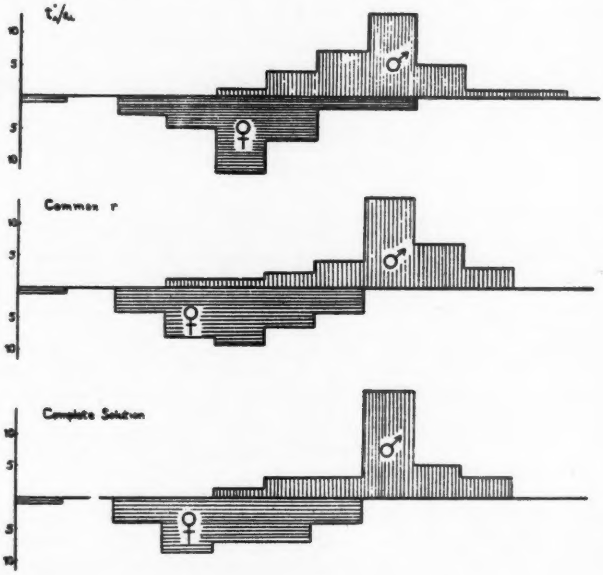
$$
\begin{aligned}
a_1 &= +2.6344 \\
a_2 &= -1.0493 \\
a_3 &= +2.4054 \\
a_4 &= -1.5983\,.
\end{aligned}
$$

For the three sets of discriminant coefficients discussed, the distribution of the resulting discriminant values is shown in Figure 2. To aid comparison, the mean value, $(y_1 + y_2)/2$, i.e., the point halfway between the means for males and females, respectively, has been made the same as has also the value of S, i.e., the sum of squares within groups. It will be seen how the two constituent distributions separate as the discriminant function is improved and also how the constituent distributions become compact, particularly in comparison with the corresponding result for Test 3 in Figure 1. It would, of course, be unnecessary to calculate the actual discriminant values, used to obtain Figure 2, in many and probably most problems that would be treated.

4. *Summary*

The weighted sum of observations on an object, so that the distinction between two groups to which such objects may have been assigned is great, has been found variously. By assuming that the correlation between various observations (as between various psy-

FIGURE 2
The Discrimination Obtained by Various Methods of Calculating the Coefficients



chological tests) is common, results have been obtained that permit two groups to be distinguished readily and effectively; the discrimination approaches closely the maximum possible with the data given in these problems. A still more simple weighting, involving simply differences between means for the two groups on the various scores, is also very satisfactory.

REFERENCES

Fisher, R. A. The use of multiple measurements in taxonomic problems. *Ann. Eugenics*, 1937, 7, 179-188.

Jackson, R. W. B. Approximate multiple regression weights. *J. exp. Education*, 1943, 11, 221-225.

Travers, R. M. W. The use of a discriminant function in the treatment of psychological group differences. *Psychometrika*, 1939, 4, 25-32.

PSYCHOMETRIKA—VOL. 10, NO. 3
SEPTEMBER, 1945

THE USABILITY OF THE CONCEPT OF "PREJUDICE"

HENRY S. DYER
HARVARD UNIVERSITY

For the purpose of determining whether the trait concept of "prejudice" is usable in the communication of meaning, representative samples of the responses of 101 ninth-grade children were submitted to a diverse group of 20 judges who were requested to rank 11 series of the responses in accordance with the amount of prejudice they were judged to exhibit. The usability, or use-value, of the concept is conceived as the extent to which the judges can agree in their ratings and is expressed in terms of the average intercorrelation of such ratings. It is shown that a null hypothesis of no use-value($\bar{r} = 0$) is untenable. The data further suggest that the concept of "prejudice" tends to have its highest use-value in situations where the factor of prejudice is commonly considered to be a matter of serious social concern.

Traits of human behavior are concepts belonging to the observers of behavior. In the communication of meaning, the usability of any such concept is a function of the extent to which individuals agree about its behavior referents. Both the number of individuals in agreement and the number of referents which are common to each individual are determining factors.

The use-value of certain trait concepts, such as mathematical ability and finger dexterity, may be assumed to be high. There is, however, a class of trait concepts for which such an assumption is not possible and about which some knowledge of their use-value seems crucial for any fruitful study of the behavior in which they are involved. One member of this class is the concept of prejudice. In view of the apparent interest in the mitigation of prejudice, it may be of some importance to determine to what extent observers can agree on what it is. In other words, can an unselected group of observers agree that certain kinds of behavior are prejudiced and certain other kinds are unprejudiced, or will the prejudices of the observers themselves preclude any significant amount of agreement? The null hypothesis of the present study is that the concept of prejudice has no use-value.

Since it is not normally possible in a controlled manner to confront an unselected group of observers with the actual behavior of an unselected group of individuals, records of such behavior were used.

TABLE 1

Samples of Responses Judged Extremely Prejudiced*

	Nature of Question	Response	Average Rank†
1.	On the treatment of Germany after the war.	If the Allies defeat Germany, I'd like to see it wiped off the map. It seems to start all the wars. Personally, I don't like Germany.	(1.7)
2.	On the political party one is most likely and least likely to join.	I'll join the Republican Party because my father and mother are Republicans. I won't join the Democratic Party. I don't know why. I just don't like it.	(4.8)
3.	On the desirability of putting boys in the same classes as girls.	(A boy.) No, I don't think I'd put boys and girls in the same classes. You can't study so well. Girls talk too much. They do more than half the talking—more than their share.	(3.9)
4.	On the desirability of putting different races and nationalities in the same classes.	I'd put different nationalities in the same classes, but not different colors. People in this country don't like the colored people. It's not a real country for them.	(3.3)
5.	On joining the labor unions.	No, I wouldn't want any employees of mine to join a labor union. I don't think very much of labor unions. I don't know why. I think it best not to have them. I wouldn't join a labor union myself. I guess I'm just against them.	(2.0)
6.	On jobs most to be preferred and least to be preferred.	That's quite a problem. There are a lot of jobs. I'd want a boss's job and be over men anyhow, so I could run the men and tell them what to do. The last job in the world I'd want would be one on the W.P.A., I guess. I wouldn't want pick and shovel and jobs like that.	(4.5)
7.	On having friends twice or half as rich as oneself.	I'd rather have my friends half as rich as me. These rich people are conceited. They're high hat. They send their children out all slicked up which more or less disgraces *us*.	(1.9)
8.	On the handling of the unemployed.	I don't know anyone with an unemployed father. They ought to take the foreigners out of jobs and only let just so many in. Americans should get the jobs first.	(2.7)
9.	On getting along with people of religions other than one's own.	I am a Congregationalist. I know Catholics, Episcopalians, Unitarians. I get along all right with everybody except Catholics.	(2.1)
10.	On the fairness of teachers.	No, teachers aren't fair. Teachers have pets. If you answer in class and someone else answers, they will try to consider the other's answer better and give them higher marks. I don't believe any students talk over their personal troubles with a teacher. I don't. She might get a bad impression of you.	(2.3)

* The nature of certain questions and responses "dates" the study. All data were gathered in the early spring of 1940.

† "Average rank" is the mean of the ranks assigned by all twenty judges. A response ranked "1" would be judged most prejudiced; a response ranked "20" would be judged least prejudiced.

Employing a standard set of ten questions, six trained persons in private interviews obtained responses from 101 ninth-grade children and recorded these responses verbatim.* The questions were constructed according to the following criteria:

(1) They should conceal the true purpose of the interview, which was to elicit responses that might be judged prejudiced or the reverse.

(2) They should be so framed as to encourage a "free" response, that is, their purpose was to get the child to talk freely about each issue in question.

(3) They should be so framed as to preclude any suggestion of a bias.

The children interviewed were representative of a considerable range in chronological age, mental age, and socio-economic status. The nationality of their parents was also diverse, although not representative of ninth-grade children in general. In the opinions of the interviewers the responses were in most cases sincere.

From the responses to each question, except one, twenty were selected as most representative of the kinds of responses obtained on the issue in question. The nature of the third question (attitude toward the opposite sex) required that forty responses be selected—twenty of the boys' responses and twenty of the girls' responses. The 220 selected responses were then presented to twenty judges who were asked to rank the responses in each set according to the degree of prejudice deemed to be exhibited. Table 1 shows the nature of each of the ten questions and provides a sample of the responses judged extremely prejudiced.

The group of judges was so small that it could hardly be considered representative of society in general; it was nevertheless highly diverse. The extent of this diversity is shown in Table 2. It is at least arguable that the use-value of the concept of prejudice for this group would be suggestive of its general use-value.

The index of use-value employed was simply the average of the intercorrelations among the ranks assigned by the twenty judges to each set of responses. Kelley† has shown that with ranked data of this sort, the average of the intercorrelations is given by the formula:

$$\bar{r} = 1 - \frac{a(4N+2)}{(a-1)(N-1)} + \frac{12\sum S^2}{a(a-1)\,N(N^2-1)}, \qquad (1)$$

* For a complete presentation of the data used in this study and the methods by which they were gathered see Dyer, H. S., Observable evidence of prejudice in ninth grade children, an unpublished dissertation on file at the Harvard University Library, Cambridge Mass., 1941.

† Kelley, T. L. Statistical method. New York: MacMillan, 1924, pp. 217-221.

TABLE 2
Summary of Information Concerning the Judges

Sex	N
Men	10
Women	10
Residence	
New England	11
The Middle West	5
The South	1
New York	1
Pennsylvania	1
China	1
Marital Status	
Married	13
Unmarried	6
Unknown	1
Ages	
60-69	1
50-59	4
40-49	3
30-39	4
20-29	8
Occupation	
Education	5
Office Worker	6
Student	2
Housewife	3
Business Executive	1
Lodging House Keeper	1
Singer	1
Lawyer	1

Religion	N
Protestant	11
Catholic	2
Hebrew	3
Spiritualist	1
None	1
Unknown	2
Nationality of Descent	
"American"	6
German	4
Polish	1
Irish	1
Swedish	1
Scotch	1
Hungarian	1
Portuguese	1
Chinese	1
Unknown	3
Political Orientation	
Radical	3
Liberal	6
Conservative	6
Unknown	5
Parenthood	
Number of Children	
3	1
2	3
1	2
0	12
Unknown	2
Education	
College	14
High School	4
Unknown	2

where $\bar{r}$ is the average of the intercorrelations, a is the number of judges, N is the number of responses judged, and S is the sum of the ranks for a given response. The formula assumes that the means and standard deviations of the ranks are equal. The data of the present study departed from this assumption in certain instances, but a test of the most drastic departure revealed a difference of only .0027 between the correct $\bar{r}$ and the $\bar{r}$ given by the formula. This difference was considered small enough to be negligible. All of the reported $\bar{r}$'s have been computed by means of Formula (1). The standard error of each $\bar{r}$ was obtained from the formula:

$$\sigma_{\bar{r}} = \frac{1-\bar{r}^2}{\sqrt{N-1}\,\sqrt{n-1}}, \qquad (2)$$

where N is the number of responses and n is the number of judges.

This is the usual formula for the standard error of r multiplied by the factor, $\frac{1}{\sqrt{n-1}}$. This reduction factor is justified on the ground that with n independent variables entering into $\bar{r}$, there are $n-1$ degrees of freedom with which to divide $\sigma_r{}^2$ in order to secure $\sigma_{\bar{r}}{}^2$.

Table 3 gives the average intercorrelation among the ranks assigned to each of the eleven sets of responses and the standard error of each $\bar{r}$.

The lowest value of $\bar{r}$ in Table 3 is .36, which is approximately eight times its standard error. The inference therefore appears warranted that chance is an unlikely cause of this lowest $\bar{r}$ value and hence of the values above it.*

TABLE 3

The Averages of the Intercorrelations Among the Ranks Assigned to Eleven Sets of Responses by Twenty Judges

	Nature of Question	$\bar{r}$	$\sigma_{\bar{r}}$
1.	Postwar treatment of Germany	.59	.034
2.	Political party preference and aversion	.49	.040
3a.	Boys' attitude toward girls	.42	.043
3b.	Girls' attitude toward boys	.40	.044
4.	On segregating races and nationalities	.59	.034
5.	On joining labor unions	.64	.031
6.	Attitude toward occupations	.36	.046
7.	Attitude toward rich and poor	.64	.031
8.	Handling of the unemployed	.51	.039
9.	Attitude toward religious groups	.46	.042
10.	On the fairness of teachers	.54	.037

With respect to their ratability as prejudiced, the eleven sets of responses appear to fall into four groups. Table 4 suggests these groupings. With the possible exception of the responses to Question 9 (attitude toward religious groups), the ratability of the responses appears to be roughly in accord with the relative importance of the issues in which prejudice is likely to be a disturbing factor in the life of society.

In view of these findings, the null hypothesis, that the concept of prejudice has no use-value, is hardly tenable. The limitations of

* A more nearly exact test of significance would be obtained by transforming the individual r's to Fisher's z, averaging the z values to secure $\bar{z}$, and then computing $\sigma_{\bar{z}}$. Since the individual r's were not readily available, this method was not feasible. However, it seems reasonable to assume that the present method is not so faulty as to invalidate the inference. (See Fisher, R. A. Statistical methods for research workers. London, 1938, pp. 202-215.)

TABLE 4

Sets of Responses Arranged According to Ratability

	Nature of Question	$\bar{r}$
5	On joining labor unions	.64
7.	Attitude toward rich and poor	.64
1.	Postwar treatment of Germany	.59
4.	On segregating races and nationalities	.59
10.	On the fairness of teachers	.54
8.	On the handling of the unemployed	.51
2.	Political party preference and aversion	.49
9.	Attitude toward religious groups	.46
3a.	Boys' attitude toward girls	.42
3b.	Girls' attitude toward boys	.40
6.	Attitude toward occupations	.36

the data permit no further positive inference. Nevertheless, the results suggest several possibilities: (1) that the concept of prejudice may have varying use-values depending upon the kinds of situations in which the behavior is observed; (2) that the use-value of the concept may be highest in just those situations where prejudice is likely to be a serious factor in social life; (3) that the use-value of the concept may be sufficiently high to permit its use as a basis for studying behavior arising in situations where prejudice is commonly considered important.

PSYCHOMETRIKA—VOL. 10, NO. 3
SEPTEMBER, 1945

A GRAPHICAL TEST FOR THE SIGNIFICANCE OF DIFFERENCES BETWEEN FREQUENCIES FROM DIFFERENT SAMPLES*

DONALD W. FISKE, LIEUTENANT, USNR
AND
JACK W. DUNLAP, COMMANDER, USNR

For testing the significance of differences between frequencies from different samples, an ellipse can easily be constructed on the basis of a formula developed on the assumption that both observed samples are random samples from the same parent population and that the best estimate of the true proportion is the weighted mean proportion of the two samples. The ellipse provides a very rapid method for testing pairs of frequencies.

In many experimental problems the investigator is concerned with the differences between frequencies of occurrence within two samples. These frequencies are expressed as proportions and the difference between the two proportions subjected to a statistical test to determine whether or not it is significant. In most experimental work no knowledge is available as to the theoretical proportions to be expected in either sample. Therefore the best estimate of the population variance, assuming the null hypothesis is to be tested, is the variance of the mean observed proportion. The mean observed proportion is defined as the weighted proportion, where the frequencies in each sample are weighted by the size of the sample (7).

It often happens that a large number of differences is to be studied for the same samples, for example in item analysis where the same items are given to two groups and it is desired to know whether the groups exhibit differential responses. In this case it is laborious to solve and re-solve the formula for the standard error of a difference, and a labor-saving device is needed. A method is outlined below for the construction of an ellipse which will enable the experimenter to tell at a glance whether or not the differences between two populations may be considered as real. Experience has shown that such an ellipse can be constructed in an hour or two, and that approximately 200 pairs of observations can be tested per hour.

* The opinions expressed in this paper are those of the authors and are not to be construed as those of the Navy Department.

Other workers have developed short-cut methods for determining the significance of differences. Nomographs for facilitating the computations are supplied by Dunlap and Kurtz (4) and by Zubin (10). Edgerton and Paterson (5) and Daniel (3) have published facilitating tables. Mosier and McQuitty (6) have developed an abac for determining the critical ratio, which can be used for any pair of equal-sized groups. Burr and Hobson (2) show a method which greatly reduces the amount of work in checking the significance of differences. Zubin (9) and Bolles and Zubin (1) have developed a method for constructing an ellipse similar to the one described below. Most of this work, however, is based on the usual assumption that the observed proportions are the true ones and that the hypothesis to be checked is that the difference between them is not due to fluctuations in sampling. The formula developed below is based upon the sounder assumption that both of the observed samples can be regarded as random samples from the same parent population and therefore the best estimate of the true proportion is the weighted mean proportion of the two samples.

A general formula for the construction of an ellipse is presented and the solution of a special case is given as an example. Examination of the general formula will indicate to the experimenter the wisdom of choosing samples of the same size in terms of expediency of computation, although equal *N*'s are not required. In order to expedite plotting the ellipse and to allow direct interpretation of frequencies, the formula has been expressed in terms of frequencies of the samples.

The basic formula for testing the significance of a difference between proportions is

$$CR = \frac{p_1 - p_2}{\sigma_{(p_1 - p_2)}} = \frac{p_1 - p_2}{\sqrt{\sigma_{p_1}{}^2 + \sigma_{p_2}{}^2}}. \tag{1}$$

Since the theoretical expectancies for p_1 and p_2 are not known and since the safest assumption that can be made is that the samples are drawn from the same population, then the best estimate of the variance of p_1 and p_2 is given by the variance of p_0.

$\sigma_{p_0}{}^2 = \dfrac{p_0 q_0}{n_1}$ and $\sigma_{p_0}{}^2 = \dfrac{p_0 q_0}{n_2}$ for samples 1 and 2, respectively,

where

$$p_0 = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \tag{2}$$

and $q_0 = 1 - p_0$.

Substituting σ_{p_0} for σ_{p_1} and σ_{p_2} in formula (1) and squaring, we may write

$$\overline{CR}^2 = \frac{(p_1 - p_2)^2}{\sigma_{p_0}^2 + \sigma_{p_0}^2} = \frac{(p_1 - p_2)^2}{\dfrac{p_0 q_0}{n_1} + \dfrac{p_0 q_0}{n_2}} \tag{3}$$

and

$$\overline{CR}^2 = \frac{(p_1 - p_2)^2}{\dfrac{p_0 q_0 (n_1 + n_2)}{n_1 n_2}}$$

or

$$\overline{CR}^2 (n_1 + n_2)\ (p_0 - p_0^2) = n_1 n_2 (p_1 - p_2)^2. \tag{4}$$

Substituting for p_0 the expression in (2) gives us

$$\begin{aligned} &\overline{CR}^2 (n_1 + n_2) \left(\frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} - \frac{(n_1 p_1 + n_2 p_2)^2}{(n_1 + n_2)^2} \right) \\ &= n_1 n_2 p_1^2 - 2 n_1 n_2 p_1 p_2 + n_1 n_2 p_2^2 . \end{aligned} \tag{5}$$

Multiplying through by $(n_1 + n_2)$ and substituting f, the frequency in a sample, for np, we may write

$$\begin{aligned} &f_1^2 \left(\overline{CR}^2 + n_2 + \frac{n_2^2}{n_1} \right) - f_1 \overline{CR}^2 (n_1 + n_2) + 2 f_1 f_2 (\overline{CR}^2 - n_1 - n_2) \\ &- f_2 \overline{CR}^2 (n_1 + n_2) + f_2^2 \left(\overline{CR}^2 + n_1 + \frac{n_1^2}{n_2} \right) = 0. \end{aligned} \tag{6}$$

Formula (6) is the general expression for the surface of the desired ellipse.

The experimenter is at liberty to select any given degree of rigor for his test of significance. Often the 5% level is chosen, and this corresponds closely to a *CR* of 2.0. Likewise the sizes of the samples are at his choice. The general formula can be solved more easily if n_1 is chosen to equal n_2.

For the purposes of illustration an ellipse is presented based on a *CR* of 2, $n_1 = 100$ and $n_2 = 100$. Substituting in the general formula gives

$$204 f_1^2 - 800 f_1 - 392 f_1 f_2 - 800 f_2 + 204 f_2^2 = 0 .$$

Dividing by 4 gives

$$51 f_1^2 - 200 f_1 - 98 f_1 f_2 - 200 f_2 + 51 f_2^2 = 0 \ , \tag{7}$$

the formula for the special case.

It is now necessary only to substitute a given value for f_2 and solve the resulting quadratic equation for two roots, which are points on the ellipse which bounds the surface within which no significant differences between frequencies are found. Note that the ordinates of the ellipse are plotted directly in terms of f_1 and f_2, eliminating the necessity for converting frequencies to proportions.

Normally five or six pairs of points are sufficient for determining the ellipse. Any convenient frequencies can be substituted, preferably those approximating the following proportions: 1.00, .95, .90, .80, .60, .40, .20, .10, .05, and .00. (Since $\sigma_p = \sigma_q$, the values for .00, .05, .10, .20, and .40 are the complements of those for 1.00, .95, .90, .80, and .60, respectively, and are computed solely as checks on the computations.) It will be apparent that if $n_1 = n_2$, then each pair of values for f_1 and f_2 can be interchanged, giving another set of points which will help to determine the ellipse more precisely.

To continue the illustration, let $f_2 = 20$, then (7) becomes

$$51f_1^2 - 2160f_1 + 16400 = 0$$

and

$$f_1^2 - 42.3529f_1 + 321.5686 = 0\,.$$

Solving the quadratic by completing the square* gives

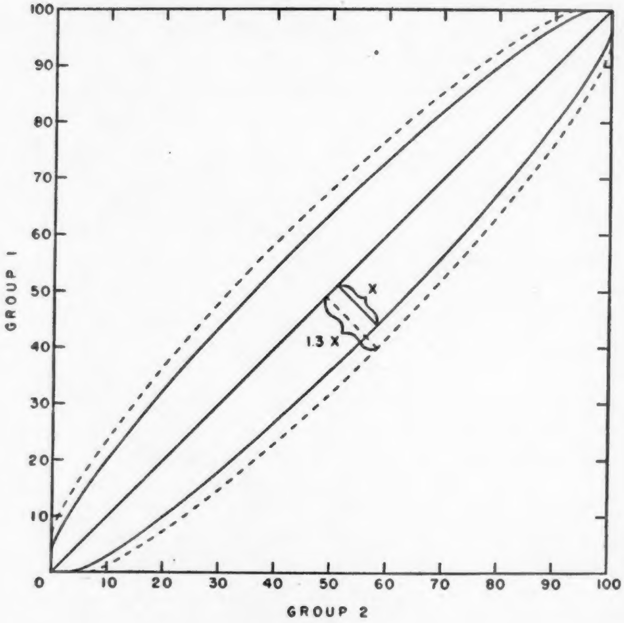
$$(f_1 - 21.1764)^2 = 448.4399 - 321.5686$$

and

$$f_1 = 32.5 \text{ and } 9.9\,.$$

When the several pairs of values have been computed, the ellipse can be plotted as in the graph below. Since the distance (x on the graph) from the straight line where $p_1 = p_2$ to the curves corresponds to a *CR* of 2.00, the approximate position of the curves representing the 1% level ($CR = 2.58$) may be determined by locating a series of points $\dfrac{2.58}{2.00}$ or roughly 1.3 times the distance from the center line to the 5% curves (1.3x on Graph 1). Other levels of significance may be approximated in the same way, or, if more precise determinations are desired, the general formula (6) may be used.

* In these computations, all decimals should be carried to at least four places since it is necessary to take the square root of the decimal value.



REFERENCES

1. Bolles, M. M., and Zubin, J. A graphic method for evaluating differences between frequencies. *J. appl. Psychol.*, 1939, 23, 440-449.
2. Burr, I. W., and Hobson, R. L. Significance of differences in proportions with constant sample frequencies in each pair. *J. educ. Psychol.*, 1943, 34, 307-312.
3. Daniel, C. Statistically significant differences in observed per cents. *J. appl. Psychol.*, 1940, 24, 826-830.
4. Dunlap, J. W., and Kurtz, A. K. *Handbook of statistical nomographs, tables, and formulas.* Yonkers-on-Hudson: World Book Co., 1932.
5. Edgerton, H. A., and Paterson, D. G. Table of standard errors and probable errors of percentages for varying numbers of cases. *J. appl. Psychol.*, 1926, 10, 378-391.
6. Mosier, C. I., and McQuitty, J. V. Methods of item validation and abacs for item-test correlation and critical ratio of upper-lower difference. *Psychometrika*, 1940, 5, 57-65.
7. Yule, G. U., and Kendall, M. G. *An introduction to the theory of statistics.* London: Charles Griffin & Co., Ltd., 1937, p. 360, paragraph 19.28(a).
8. Zubin, J. Note on a transformation function for proportions and percentages. *J. appl. Psychol.*, 1935, 19, 213-220.
9. Zubin, J. Note on a graphic method for determining the significance of the difference between group frequencies. *J. educ. Psychol.*, 1936, 27, 431-444.
10. Zubin, J. Nomographs for determining the significance of the differences between the frequencies of events in two contrasted series or groups. *J. Amer. statist. Ass.*, 1939, 34, 539-544.